Universidade de Aveiro Departamento de Química
2014

Ana Carolina Marianito Serrano

Análise de genomas humanos de indivíduos portugueses

Universidade de Aveiro Departamento de Química
2014

Ana Carolina Marianito Serrano

Análise de genomas humanos de indivíduos portugueses

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Bioquímica, ramo Métodos Biomoleculares, realizada sob a orientação científica da Doutora Conceição Egas, responsável pela unidade de Serviços Avançados do Biocant e do Professor Doutor Manuel Santos, professor associado do Departamento de Biologia da Universidade de Aveiro.

biocant
CENTRO DE INOVAÇÃO EM BIOTECNOLOGIA

o júri

Presidente

Prof. Doutor Bruno Miguel Rodrigues das Neves
Professor auxiliar do Departamento de Química da Universidade de Aveiro

Doutora Maria Conceição Venâncio Egas
Diretora da Unidade de Sequenciação Avançada do Biocant

Doutor Nuno André Rodrigues Faustino
Investigador na GenePredit

agradecimentos

A realização desta tese não teria sido possível sem a colaboração, presença e apoio de diversas pessoas a quem gostaria de deixar expresso o meu sincero agradecimento.

À minha orientadora, Doutora Conceição Egas, pela oportunidade, apoio, orientação e ensinamentos que permitiram o desenvolvimento deste trabalho.

Ao Professor Doutor Manuel Santos, por ter aceitado coorientar esta tese.

Aos meus colegas de unidade do Biocant, por me terem acompanhado ao longo desta jornada, pelo apoio, pela motivação e boa disposição que sempre me disponibilizaram.

À Susana Carmona pela forma altruísta com que me ajudou, todos os dias e a qualquer hora, pelo apoio, ânimo e amizade que foram essenciais.

Aos meus colegas da Universidade de Aveiro, por me terem recebido e permitido aprender com eles, em especial à Maria Quental pela amizade e ao Filipe Rodrigues pelo companheirismo e lealdade.

À Ana Carvalho, por ter marcado tão positivamente a minha passagem por Aveiro.

Aos meus colegas de direção da Associação Portuguesa dos Técnicos de Análises Clínicas e Saúde Pública, pelo apoio e compreensão demonstrados nos momentos de ausência forçada por este trabalho.

Aos meus amigos, os de sempre, por compreenderem os meus períodos de ausência e por caminharem comigo no alcance dos meus objetivos.

À Joana Matos e ao Pedro Canhão, por estarem comigo todos os dias e serem fundamentais.

Ao Armando Caseiro, o meu eterno mestre, por tudo o que me ensinou e continua a ensinar, por ter sido a força motriz deste trabalho e por me fazer feliz.

À minha sobrinha Maria Serrano, pelo brilho do seu sorriso que cria em mim a coragem e a força para tentar ser sempre mais e melhor.

Aos meus pais, Gilberto e Augusta, e ao meu irmão Pedro, os melhores, por serem a quem devo tudo e a quem pertencem todas as minhas vitórias, eles que são a minha certeza, a minha força e o meu maior exemplo!

palavras-chave

Sequenciação; *Next-generation Sequencing* (NGS); Genoma humano; Variante; MicroRNA; Ancestralidade.

resumo

O advento da tecnologia de sequenciação de última geração (NGS) possibilitou a obtenção de um conjunto de genomas individuais, graças ao aumento significativo na capacidade de produção de dados e à consequente redução de custo e tempo associado. Ultrapassada a dificuldade de ler o genoma, o desafio passa agora por compreendê-lo. Este trabalho apresenta a análise dos primeiros quatro genomas de indivíduos portugueses sequenciados com elevada cobertura (≈ 34x), pela tecnologia Illumina. Da análise resultou a identificação, em média, de 4.586.683 variantes em cada genoma, correspondendo 4.146.858 a polimorfismos de um único nucleótido (SNPs) e 439.825 a pequenas inserções e deleções (INDELs). Na região codificante dos quatro genomas foram identificados, em média, 24.791 SNPs, dos quais 12.139 foram alterações *missense*, 140 foram alterações *nonsense* e 48 foram SNPs conducentes à perda de codão *stop*. O impacto funcional destas alterações foi previsto, tendo-se identificado uma média de 149 variantes por genoma com potencial efeito deletério. A análise estendeu-se à região não codificante, tendo sido identificados, em média, 52 miRNAs com variantes por genoma. Na região intrónica, ≈1.008.170 de SNPs reguladores (rSNPs) da expressão génica, ou SNPs em *linkage disequilibrium* com rSNPs, foram identificados. A análise da variabilidade em cada genoma possibilitou ainda a determinação da ancestralidade genética dos indivíduos portugueses, a qual denotou uma diferenciação das restantes populações embora num posicionamento próximo das caucasianas. Os nossos resultados sugerem que a análise global da variabilidade genética faculta uma poderosa oportunidade para identificar características individuais, e/ou populacionais, biologicamente significativas e/ou clinicamente úteis.

keywords

Sequencing; Next-generation Sequencing (NGS); Human genome; Variation; MicroRNA; Ancestrality.

abstract

The advent of next-generation sequencing technologies enables the complete sequencing of individual genomes, due to the significant increase in data production and the decrease in time and costs associated. Now that the genome sequence is available, the challenge is to interpret the information encoded. This study analyzed the first four human genomes from Portuguese individuals, sequenced with high coverage (≈ 34x) with Illumina technology. In average 4.586.683 variants were identified in each genome, corresponding to 4.146.858 single nucleotide polymorphisms (SNPs) and 439.825 small insertions and deletions (INDELs). In the coding region, an average of 24.791 SNPs was identified, of which 12.139 were missense, 140 nonsense and 48 loss of stop codons. The variants functional impact was predicted and, in average, 149 variants were identified as potentially deleterious per genome. The analysis also targeted noncoding region, where in average 52 miRNAs were altered per genome. In the intronic region ≈1.008.170 gene expression regulatory SNPs (rSNPs) or SNPs in linkage disequilibrium with rSNPs were identified. The genetic ancestrality of each genome was determined from its variants. The Portuguese genomes clustered close to the caucasian populations. Our results suggest that the global genetic variability analysis is a powerful opportunity to identify individual and/or population characteristics with biologic or clinical relevance.

I - Índice

II - Lista de figuras

III - Lista de tabelas

IV - Lista de siglas e abreviaturas

Ago – Argonaute

AIMs – Marcadores informativos de ancestralidade

APS – Persulfato de amónia

ATP – Adenosina trifosfato

BAM – *Binary alignment map*

BCAAs – Catabolismo dos aminoácidos de cadeia ramificada

BCKD – Desidrogenase dos cetoácidos de cadeia ramificada

C. elegans – Caenorhabditis elegans

CNVs – *Copy number variants*

dbSNP – *The Database for Single Nucleotide Polymorphism*

ddNTPs – didesoxinucleótidos trifosfato

DGCR8 – *DiGeorge critical region 8*

DNA – Ácido Desoxirribonucleico

dNTPs – Desoxinucleótidos trifosfato

EDTA – Ácido etilenodiaminotetracético

ENCODE – Enciclopédia de Elementos do DNA

ESP – Projeto de sequenciação de exomas

F_{ST} – Índice de fixação de *Wright*

GATK – *The Genome Analysis Toolkit*

Gb – Gigabase

GFF3 – *Generic Feature Format 3*

GO – *Gene Ontology*

GRC – *Genome Reference Consortium*

GWAS – *Genome-wide association studies*

HGDP – *Human Genome Diversity Project*

HGMD – *Human Genome Mutation Database*

HIF – Fator indutor de hipóxia

$\overline{H_S}$ – Média da heterozigotia

H_T – Heterozigotia total

IGV - *Integrative Genomic Viewer*

INDELs – Inserções/deleções

Kb – Quilobase

LASER – *Locating Ancestry from Sequence Reads*

LD – *Linkage Disequilibrium*

LRT - *Likelihood Ratio Test*

MAF – Frequência do alelo minor

Mb – Megabase

mtDNA – DNA mitocondrial

miRISC – complexo ribonucleico de silenciamento

miRNA – microRNA

mRNA – RNA mensageiro

NCBI – *National Center for Biotechnology Information*

NF-kB – Fator nuclear kappa B

NGS – *Next-Generation Sequencing*

NHGKI – *National Genome Research Institute*

NHLBI – *National Heart, Lung and Blood Institute*

OMIM – *Online Mendelian Inheritance in Man*

pb – pares de bases

P bodies – Processing bodies

PC1 – Primeira componente principal

PC2 – Segunda componente principal

PCA – Análise dos componentes principais

PCR – reação em cadeia da polimerase

PGH – Projeto do Genoma Humano

PHANTER – *Protein Analysis Through Evolutionary Relationships*

Phylop – *Phylogenetic p-Value*

PolyPhen – *Polymorphism Phenotyping*

PharmaGKB – *Pharmacogenomics Knowledge Base*

PTP – *Picotaterplate*

qRT-PCR – Reação em cadeia da polimerase quantitativa após transcrição reversa

RNA – Ácido ribonucleico

rSNP - Polimorfismos de um único nucleótido regulador

SAM – *Sequence alignment map*

SIFT – *Sorting Intolerant from Tolerant algoritm*

SNPs – Polimorfismos de um único nucleótido

STRING – *Search Tool for the Retrieval of Interaction Genes/Proteins*

STRs – *Short tandem repeats*

SVs – Variantes estruturais

Ts – Transições

Tv - Transversões

UCSC – *University of California Santa Cruz*

VCF – *Variant call format*

ZMW – *Zero-mode waveguides*

Introdução

1. Sequenciação do genoma humano

1.1. Do início da era genómica ao Projeto do Genoma Humano

Em 1953, a publicação que sugeriu a estrutura do ácido desoxirribonucleico (DNA), da autoria de James Watson e Francis Crick (1), revelou-se basilar na evolução científica que hoje permite construir a história da humanidade, e possivelmente, modificar o futuro da mesma. Motivadas pelo marco então alcançado, na década de 50, surgiram áreas científicas como a bioquímica de nucleótidos e a biologia molecular, as quais viriam a estar na génese da genómica, disciplina que utiliza a tecnologia que permite sequenciar o DNA. A sequenciação de DNA é a metodologia que determina a ordem pela qual as bases nucleotídicas constituem um fragmento de DNA, e surgiu 20 anos após a publicação de Watson e Crick (2). Em 1973, Walter Gilbert e Allan Maxam publicaram a sequência de 24 pares de bases (pb) do operão *lac*, utilizando o método de sequenciação inicial, baseado na degradação das bases do DNA de um modo específico por reagentes químicos (3). Consecutivamente ao método de Maxam-Gilbert, Frederick Sanger e colaboradores publicaram em 1977 uma metodologia de sequenciação por síntese de DNA (4). Sanger e colaboradores mostraram conseguir determinar unidimensionalmente a sequência de bases do genoma, recorrendo a um método enzimático que rapidamente suplantou o método de Maxam-Gilbert (2). O método inicial de Sanger utiliza didesoxinucleótidos trifosfato (ddNTPs) radiomarcados, nucleótidos que sofreram uma substituição, na posição 3' da desoxirribose, de um grupo OH por um H. Esta modificação conduz à impossibilidade de formação de uma ligação fosfodiéster aquando da reação de síntese de uma cadeia de DNA, terminando a síntese sempre que um ddNTP é adicionado. A sequenciação ocorre numa série de 4 reações em separado (A, C, G e T), onde é adicionada uma quantidade de ddNTPs às reações (ddATPs, ddCTPs, ddGTPs, ddTTPs, respetivamente), em conjunto com os respetivos desoxinucleótidos trifosfato (dNTPs) convencionais. Assim, a extensão de um *primer*, promovida por uma DNA polimerase, é estocasticamente terminada pela incorporação de um ddNTP que compete com o respetivo dNTP. O produto da reação é um conjunto de cadeias oligonucleotídicas de comprimento diverso. Após a eletroforese e subsequente autoradiografia, é possível deduzir, a partir do tamanho dos fragmentos, a sequência de bases que compõem o fragmento de DNA analisado (4). Sanger veio assim

abrir portas para a era da genómica, induzindo um forte estímulo na comunidade científica para sequenciar o genoma humano em toda a sua extensão. No entanto, embora tenham surgido melhorias no método de Sanger, tais avanços não foram suficientes para tornar a abordagem de sequenciação verdadeiramente adaptada às necessidades de *high throughput* que a sequenciação de todo o genoma humano requer (2).

Um enorme salto foi conseguido com a descoberta da DNA polimerase termoestável, o que conduziu não só ao desenvolvimento da reação da polimerase em cadeia (PCR), como a uma melhoria considerável na enzimologia da sequenciação (5). A incorporação da amplificação na reação de sequenciação também motivou melhorias notórias e permitiu começar a reação com uma quantidade significativamente menor de DNA molde. Avanços na bioquímica dos nucleótidos foram também importantes, com introdução de ddNTPs marcados com corantes fluorescentes, ao invés de radioisótopos. Decorrente de todos estes avanços surgiu o método automático de sequenciação de DNA baseado na bioquímica de Sanger, onde cada ddNTP marcado com um fluorocromo diferente permite que uma única reação ocorra, sendo o produto desta sujeito a uma eletroforese capilar (6). Num determinado ponto do capilar incide um feixe laser que excita os fluorocromos e provoca a emissão de fluorescência de um determinado comprimento de onda, em função do fluorocromo encontrado. Tal abordagem permitiu eliminar um esforço manual considerável e várias fontes de erro, na mesma medida em que o rendimento de cada reação aumentou significativamente com a automatização, permitindo que em cada corrida 96 amostras fossem carregadas num gel (7). A utilização de *software* específicos permite a tradução do eletroferograma obtido em sequências de DNA, além de gerar probabilidades de erro para cada base identificada (8). Ainda que tais avanços tenham sido encorajadores, sequenciar o genoma humano significa sequenciar aproximadamente três mil milhões de bases, ou seja 3 gigabases (Gb). Atendendo ao facto de uma corrida com 96 capilares gerar apenas 48 mil bases, ou 48 quilobases (Kb), o que equivale a 0,000048 Gb, a grandeza que se impunha ao desafio é notória (9). Contudo, o constante desenvolvimento e melhoria da bioquímica de Sanger levou a que a ideia de analisar todo o genoma tenha sido proposta formalmente em 1990 (10).

O Projeto do Genoma Humano (PGH) foi programado decorrer durante 15 anos, tendo sido descrito como um esforço internacional para determinar a sequência de todas as

bases do genoma humano, identificar e mapear os 23 pares de cromossomas, armazenar a informação em bases de dados e desenvolver ferramentas que permitissem, mediante a análise dos dados obtidos, o estudo da biologia e da medicina (11). Ciente do desafio, o PGH salientou a necessidade de desenvolvimento contínuo da tecnologia, a fim de manter o projeto dentro da calendarização estabelecida. Em 1998 uma metodologia de sequenciação por *shotgun* foi proposta com a finalidade de aumentar a aquisição de dados da sequência (12). Esta abordagem consiste na quebra aleatória do DNA genómico seguindo-se a clonagem dos fragmentos em vetores antes da sequenciação, o que se traduz numa melhoria de tempo à custa da simplificação do processamento computacional. Em 2001 surge a publicação inicial do PGH, referente a cerca de 90% do genoma, com a restante percentagem não sequenciada localizada predominantemente ao nível da heterocromatina (13). No ano de 2003 foi celebrada a conclusão do PGH, tendo este sido publicado no ano seguinte, ficando o feito notavelmente marcado na história da ciência (14). A sequência final publicada continha aproximadamente 99,7% do genoma, com um erro por cada 100.000 bases. Continha apenas 300 *gaps*, cerca de 28 megabases (Mb) em regiões de eucromatina, predominantemente regiões repetidas, e cerca de 200 Mb ao nível da heterocromatina, nomeadamente os grandes centrómeros e os braços curtos dos cromossomas acrocêntricos. Foram 13 anos de percurso que permitiram passar da observação fenotípica ao conhecimento da sequência do genoma. Da combinação da bioinformática, com a bioquímica e a biologia molecular, surgiu desde então a necessidade de criar tecnologia capaz de repetir o feito num processo económico e célere com vista à pretendida meta da medicina personalizada (15).

1.2. *Next-Generation Sequencing*

Após a prerrogativa proporcionada pela conclusão do PGH, a tecnologia de sequenciação do DNA foi alvo de uma intensa transformação metodológica com vista ao efetivo avanço científico do conhecimento do genoma (16). Embora a bioquímica de Sanger tenha sido fulcral ao disponibilizar um conjunto de genomas completos de organismos modelo, tal como o genoma humano, à data da conclusão do PGH poucas vias restavam para a otimização do método tendo em vista a redução de custos e tempo (17). Assim, um desenvolvimento significativo na complexa interação de enzimologia,

bioquímica, alta-resolução ótica, *hardware* e engenharia de *software* levou ao surgimento de tecnologia de sequenciação de última geração, ou *Next-Generation Sequencing* (NGS).

As novas plataformas de sequenciação, que surgiram após o ano de 2004, utilizam uma igual estratégia metodológica, ainda que por diferentes abordagens (16). No fundo, as novas tecnologias diferem da abordagem tradicional na medida em que abdicam do processo de clonagem, utilizando o DNA a sequenciar para construir uma biblioteca genómica de fragmentos desse mesmo DNA (18). Cada tecnologia pretende amplificar cadeias simples da biblioteca de fragmentos e realizar as reações de sequenciação nas cadeias amplificadas. Para que tal suceda, a cada extremidade dos fragmentos presentes na biblioteca é adicionado um adaptador, por ação de uma DNA ligase. Estes adaptadores são sequências universais e específicas de cada plataforma, o que permite amplificar por PCR os fragmentos da biblioteca, quer seja sobre uma base sólida ou numa *bead*, que são covalentemente derivatizadas com sequências de adaptadores complementares aos dos fragmentos (18). Como resultado da amplificação obtêm-se *clusters* de fragmentos, sendo cada *cluster* originário de um único fragmento da biblioteca. Tal irá fornecer um sinal suficiente de cada fragmento na reação de sequenciação. É devido a esta abordagem combinada de amplificação com a subsequente sequenciação, que estas plataformas utilizam uma metodologia de sequenciação massiva em paralelo. Estas plataformas permitem assim sequenciar, em simultâneo, centenas de milhares a milhões de *clusters*, durante uma única corrida, demonstrando uma enorme capacidade de obtenção de dados. Esta maximização do rendimento em cada corrida traduz-se em custos consideravelmente mais baixos de sequenciação (19). Contudo, uma limitação evidente distingue as tecnologias de NGS da sequenciação convencional. Esta reside no comprimento das *reads* obtidas, ou seja o número de nucleótidos obtidos em cada fragmento a ser sequenciado. Sendo o comprimento das *reads* no método de Sanger limitado por fatores relacionados com a eletroforese (nomeadamente a percentagem de acrilamida, no gel de poliacrilamida), é possível obter um comprimento superior de *reads* comparativamente ao comprimento obtido por NGS, o qual é determinado em função da relação sinal-ruído (16). Dada a dificuldade de assemblar *reads* curtas, nomeadamente em regiões de elevado teor repetitivo, algoritmos específicos para a abordagem de assemblagem deste tipo de *reads* foram desenvolvidos (20, 21). Com efeito, nos últimos anos a tecnologia de sequenciação tem evoluído no sentido de obter *reads* de comprimento mais longo. Ainda no seguimento

das *reads* obtidas por NGS, surgem duas abordagens distintas de realizar sequenciação: *single-end sequencing* e *paired-end sequencing.* Enquanto o *single-end sequencing* utiliza uma biblioteca cujos fragmentos são sequenciados apenas de um dos lados, o *paired-end sequencing* utiliza uma biblioteca cujos fragmentos são sequenciados nos dois lados deste, a partir de cada uma das extremidades (16). Utilizando a combinação destas abordagens é possível obter na sequência uma ordem de longo alcance e orientação (através de *paired-end reads*), e montar de uma forma localizada as regiões de sequência difícil (proporcionado por *single-end reads*). A utilização destes dois tipos de *reads* tornou-se essencial quando se pretende montar um genoma pela primeira vez (sequenciação *de novo*) (22).

Recentemente uma nova geração de sequenciadores voltou a revolucionar a tecnologia. Estes novos sequenciadores denotam uma ausência de amplificação, utilizando por sua vez um sistema de deteção altamente sensível, para caracterizar diretamente a molécula individual de DNA. Deste modo, na vanguarda da tecnologia de sequenciação estão os designados sequenciadores de molécula única (16).

1.2.1. Sequenciadores de segunda geração

O principal componente, que permite aos sequenciadores de segunda geração utilizarem uma tecnologia de sequenciação massiva em paralelo, é a amplificação por PCR de uma biblioteca genómica. Apesar desta similaridade, as plataformas de segunda geração apresentam-se bastante diversificadas em termos de bioquímica de sequenciação.

A abordagem por pirosequenciação (454 Life Sciences/Roche) foi a primeira tecnologia disponível na era NGS, em 2005 (23). Esta tecnologia inicialmente envolve a geração de uma biblioteca de fragmentos de DNA, a qual é submetida a uma amplificação em massa por PCR de emulsão, sobre a superfície de centenas de milhares de *beads* de agarose. A formação de micelas aquosas permite formar micro reatores, onde cada *bead* é individualmente incorporada, juntamente com reagentes de PCR. A amplificação de um único fragmento da biblioteca genómica ligado a cada *bead*, permite formar pelo menos 1.000.000 de cópias desse fragmento por *bead*, a fim de produzir um sinal detetável na reação de sequenciação (24). Após a quebra das micelas, cada *bead* é incorporada num poço de uma placa *picotaterplate* (PTP), juntamente com reagentes de sequenciação. As

bases são adicionadas sequencialmente à placa pela mesma ordem, sendo que sempre que um nucleótido é adicionado é libertado pirofosfato, que ao reagir com persulfato de amónia (APS) origina adenosina trifosfato (ATP) e sulfato, numa reação catalisada pela enzima ATP-sulfurilase. O ATP reage com luciferina e O_2, numa reação catalisada pela enzima luciferase, que leva à emissão de luz. A luz produzida é detetada, traduzindo-se na sequência de bases nucleotídicas (23).

A deteção por fluorescência da sequenciação realizada por hibridização (SOLiD, Applied Biosystems) é outra das tecnologias de segunda geração, que surgiu no mesmo ano que a pirosequenciação (25). Assemelha-se com a tecnologia de pirosequenciação na medida em que amplifica a biblioteca genómica por meio de uma PCR de emulsão, anteriormente descrita (24). As *beads* obtidas são depositadas numa placa de vidro onde a sequenciação é realizada por múltiplas reações de hibridização de uma população parcialmente degenerada de octâmeros, catalisadas por uma DNA ligase. Cada octâmero está estruturado de forma a ter uma marcação de fluorescência num nucleótido específico, sendo este nucleótido clivado quimicamente após a ligação e deteção do sinal. O registo das fluorescências emitidas, de forma sequencial, permitem ler a sequência de bases do fragmento (25).

Em 2007, surge uma metodologia que utiliza terminadores reversíveis marcados com fluorocromos (Illumina, Solexa Technologies) (26). Esta abordagem utiliza uma amplificação em ponte dos fragmentos de DNA da biblioteca genómica, em detrimento da PCR de emulsão. Neste método, as sequências adaptadoras são ligadas aos fragmentos de DNA, os quais são ligados a uma superfície sólida. Nesta, as moléculas são amplificadas em *clusters*, sendo cada um composto por cerca de 1.000 cópias do fragmento (27). A sequenciação massiva em paralelo ocorre por adição de terminadores reversíveis, nos quais a presença de um radical quimicamente clivável na posição 3'OH da desoxirribose permite a incorporação de uma única base em cada ciclo, detetando um de quatros fluorocromos, também eles quimicamente cliváveis, correspondendo à identificação de cada nucleótido (26).

Uma nova metodologia surge em 2010, baseada na deteção da sequência por alterações de pH (Ion Torrent/Proton, Life Techonologies) (28). Nesta abordagem, à semelhança de outras, a biblioteca de fragmentos genómicos é amplificada por PCR de

emulsão (24). Contudo, as *beads* são colocadas num *chip* iónico de silício semicondutor, que consiste num sequenciador acoplado a um sensor iónico para detetar variações de pH dentro de poços individuais que contêm uma única *bead* e onde a reação ocorre. Quando os nucleótidos são incorporados na cadeia, um protão (H^+) é libertado como subproduto da incorporação. A carga iónica é então registada pelo detetor de silício, alterando a voltagem de modo a que cada pico de voltagem corresponda a um evento de incorporação. Esta metodologia implica a incorporação de nucleótidos por uma ordem sistemática, uma vez que estes não possuem nenhum tipo de marcação que os identifique (28).

1.2.2. Sequenciadores de terceira geração

O advento de plataformas de segunda geração foi decisivo no avanço revolucionário da tecnologia de sequenciação do DNA. Todavia, dificuldades associadas à sequenciação de uma população de fragmentos amplificados por PCR tornaram-se críticas e impuseram a necessidade de resolução. Nomeadamente, a ocorrência de erros pela polimerase durante a construção da biblioteca genómica, que podem mimetizar bases variantes no genoma original; bem como a amplificação preferencial de certos fragmentos na biblioteca, que confere uma discrepância em termos de abundância final de fragmentos (16). Torna-se claro, portanto, que o desenvolvimento de uma tecnologia com sensibilidade para uma única molécula de DNA seria a resolução de algumas limitações, evoluindo-se deste modo para uma terceira geração de plataformas. Sensores com uma sensibilidade sem precedentes, permitiram desenvolver sistemas miniaturizados à escala molecular, dispensando a fase inicial de amplificação e, consequentemente, os erros associados. Esta miniaturização à escala nanoscópica tem ainda como vantagem a maior rapidez e economia de processos, encontrando-se já várias plataformas disponíveis (revisto em (29)).

Uma das plataformas para a sequenciação de uma única molécula de DNA serve-se da tecnologia de nanoporos (Oxford Nanopore Technologies) (30). Nesta abordagem revolucionária, a sequência de nucleótidos da molécula de DNA é conduzida através de nanoporos, semelhantes a uma proteína de membrana. Flutuações na condutância promovidas pela passagem da molécula de DNA através do poro, ou a deteção de interações de cada base com o poro, são métodos utilizados para deduzir a sequência de nucleótidos da molécula de DNA.

A monitorização em tempo real da atividade da DNA polimerase é um processo comum a outras abordagens de terceira geração. Uma destas abordagens faz a ponte entre a nanotecnologia e a biologia molecular, e baseia-se na utilização de *Zero-Mode Waveguides* (ZMW) para detetar a incorporação de nucleótidos numa cadeia em formação (Single Molecule Real Time, Pacific Biosciences) (31). Durante a reação de sequenciação, um fragmento da molécula de DNA é elongado por uma DNA polimerase com dNTPs que são marcados com fluorocromos diferentes mediante as bases, na fração terminal do fosfato. Esta DNA polimerase encontra-se covalentemente ligada na base de cada ZMW, que estão presentes num *chip* onde ocorre a reação. A sequência de DNA é determinada através da fluorescência dos nucleótidos, a qual é imediatamente interrompida após a formação de uma ligação fosfodiéster que provoca a difusão do fluorocromo para fora da zona ZMW.

Embora sejam reveladoras de uma enorme evolução metodológica, existem alguns desafios colocados a este tipo de abordagens, centrando-se estes maioritariamente em torno da relação sinal-ruído. Neste sentido, a investigação prossegue no sentido de desenvolver sensores mais precisos a um nível extraordinariamente baixo de sinal produzido, bem como mecanismos que permitam ler uma única molécula várias vezes, para produzir uma *read* consenso com uma precisão extremamente elevada (16).

1.3. Genomas humanos individuais

A conclusão do PGH permitiu desafiar a Genética Humana a uma transição para a Genómica Humana e o advento das tecnologias NGS tornou esse pressuposto exequível. Atribuindo ao estudo do genoma uma realidade até então inatingível, rapidamente surgiu a intenção de sequenciar genomas individuais (32). Em 2007, Levy e colaboradores publicaram a primeira sequência de um genoma humano individual (Craig Venter), assemblado *de novo* a partir de 32 milhões de *reads*, com ≈700 pb de comprimento, obtidas pelo método de Sanger (33). A dificuldade de assemblar *de novo reads* curtas provenientes de NGS, a par com a existência de uma sequência de referência, motivaram a obtenção de genomas individuais por re-sequenciação. Esta abordagem promove o mapeamento de cada *read* a uma sequência do genoma de referência, levando ao surgimento de uma sequência consenso que é semelhante mas não necessariamente idêntica à referência. Esta sequência de referência do genoma humano é uma sequência consenso haploide, derivada de múltiplos indivíduos (34). Em 2008 surgiu o primeiro

genoma humano completo sequenciado pela tecnologia NGS (pirosequenciação), através de uma abordagem de re-sequenciação (35). A sequência genómica que protagonizou tal marco pertence a James Watson e serviu como guia para o conjunto de genomas individuais que se seguiram desde então. Ainda no ano de 2008, Bentley e colaboradores sequenciaram o genoma completo de um indivíduo nigeriano, pela primeira vez através da tecnologia NGS que utiliza terminadores reversíveis marcados com fluorocromos, seguindo-se a obtenção da primeira sequência genómica de um indivíduo asiático sequenciada pela mesma tecnologia (36, 37). Em 2009 surgiu a sequência do primeiro indivíduo coreano (38). No ano seguinte, foi sequenciado o genoma do primeiro indivíduo irlandês e desde então outros genomas de variadas populações se seguiram, pelas diversas tecnologias de sequenciação (39-43).

Não obstante do notável esforço empregue nos últimos 5 anos para a leitura da sequência de bases que definem determinados indivíduos, claro está que tal não significa a compreensão dessa leitura. Compreender o genoma humano passa por conhecer a função dos genes, das consequentes proteínas e a interligação destas em mecanismos biomoleculares. Mais, compreender os genomas humanos individuais denota a necessidade de encontrar as diferenças entre esses genomas e assimilar o seu significado. No fundo, é a procura pela variação existente no genoma (15).

1.3.1. Variação do genoma humano

O genoma humano é constituído por cerca de 3 mil milhões de pb de DNA, divididos por 46 cromossomas nucleares (44 autossomas e 2 sexuais) e 1 cromossoma mitocondrial consideravelmente menor. A atual estimativa do número de genes que constituem todo o genoma ronda os 23.000, distribuídos de forma não-aleatória pelos cromossomas, representando as zonas de codificação destes menos de 2% da totalidade genómica (44). A variação genética, que existe ao longo de toda a extensão do genoma, denota uma enorme importância funcional e como tal tem sido o alvo da análise de genomas, a vários níveis. Estima-se que dois indivíduos aleatoriamente escolhidos tenham sequências 99,9% idênticas. Sensivelmente 0,1% da variabilidade no genoma humano representa ≈3 milhões de alterações por indivíduo (45). Efetivamente, já em 1902 era estabelecido o conceito de "individualidade química", contudo só com a disponibilidade da

sequência humana completa, a caracterização dos diferentes tipos de variantes genéticas foi possível (46).

Polimorfismos de um único nucleótido (SNPs) são variantes de uma única base na sequência de DNA e representam a forma mais abundante de variação no genoma. No que à abundância diz respeito, seguem-se pequenas inserções/deleções (INDELs), variantes que são mais drásticas em relação a alterações na sequência, como na estrutura de codões e potencialmente na funcionalidade dos elementos genéticos transcritos (2). Uma fração de INDELs, que se manifesta por um número variável de repetições em *tandem*, é outra forma de variação conhecida como *short tandem repeats* (STRs). Esta forma de variação consiste numa pequena sequência de comprimento variado que é repetida por um número de vezes também ele variável (microssatélites e minisatélites). *Copy number variants* (CNVs) e outras alterações estruturais (SVs) incluem regiões genómicas de duplicações, deleções, inversões e translocações. Associadas a grandes alterações na sequência genómica, este tipo de variantes representam, por norma, um desequilíbrio no balanço biológico normal da diploidia num determinado locus, estando por isso associadas em grande parte a genomas tumorais ou com doenças raras de acentuada gravidade (33). Porque a informação das variantes é extremamente útil em vários aspetos genéticos de valor clínico e populacional, alguns projetos têm sido desenvolvidos no sentido de caracterizar variantes pontuais (SNPs e INDELs) e estruturais (CNVs e SVs) em diferentes populações (47-49). O projeto HapMap surgiu de um consórcio internacional, com vista à criação de um catálogo de variantes genéticas comuns que ocorrem nos seres humanos (50). Este visa caracterizar diferentes populações com ancestralidade africana, asiática e europeia, relativamente a variantes comuns, à sua frequência e aos seus padrões de expressão, nomeadamente a caracterização de haplótipos. O impacto do HapMap tem sido enorme e decisivo na investigação em vários campos, desde o estudo de doenças complexas, à genética de populações e à genética evolutiva (48). No encalço dos dados disponíveis, a realização de *Genome-Wide Association Studies* (GWAS) beneficiou de um enorme impulso. Os GWAS estabelecem uma relação estatística entre uma ou mais variantes de uma determinada região genómica, com a presença ou ausência da condição clínica (51). Não obstante do valor destes estudos, estes denotam uma perda de herdabilidade, na medida em que as variantes comuns (frequência do alelo minor – MAF>5%) que estes associam com determinada característica, apenas contribuem para uma pequena fração do genótipo

responsável (<10%). Deste modo, surge a necessidade de catalogar variantes de baixa frequência ou raras, que possam estar na causa de determinado estado clínico (MAF<5%) (52). Assim, surgiu em 2007 o projeto internacional dos 1000 Genomas (53). O objetivo deste é encontrar variantes genéticas com frequência de pelo menos 1% na população estudada, pela sequenciação de ≈2.500 indivíduos de mais de 20 etnias. No âmbito dos 1000 Genomas, foi publicado em 2012 o resultado da sequenciação de todo o genoma com baixa cobertura (2 – 6x) e a sequenciação de todo o exoma (50 – 100x) de 1.092 indivíduos de 14 populações diferentes do continente europeu, asiático, africano e americano. Até ao momento, o projeto encontrou 40 milhões de variantes genéticas, das quais 38 milhões são SNPs (54% destes são novos SNPs), 1,4 milhões de INDELs e 14.000 CNVs/SVs (47). Recentemente surgiu o projeto de sequenciação de exomas (ESP), do *National Heart, Lung and Blood Institute* (NHLBI), o qual conduziu à sequenciação do exoma de 6.515 indivíduos, com ancestralidades afroamericana e americana europeia, com o objetivo de catalogar apenas o conjunto de variantes que se encontram na região de codificação de proteína (54).

Embora o estudo das variantes, anteriormente descritas, seja preponderante na evolução da caracterização funcional da sequência de bases lida, um outro conjunto de características é decisivo em termos funcionais. As características epigenéticas, como o enrolamento da cromatina e a complexa variedade de histonas e proteínas não histonas, influenciam a atividade de diversos elementos genómicos (55). Estes aspetos também são explorados no estudo da variabilidade, sendo altamente dinâmicos e subjacentes ao controlo da expressão de genes e de outras sequências genómicas com profunda relevância na função celular e do organismo. Ciente de toda esta panóplia que figura no genoma e que é determinante na obtenção de estudos funcionais, surge em 2003 um consórcio internacional denominado Enciclopédia de Elementos do DNA (ENCODE), fundado pelo *National Genome Research Institute* (NHGKI) (56). O ENCODE visa descobrir elementos funcionais não codificantes no genoma humano, usando informação da cromatina; modificações nas histonas; posicionamento no nucleossoma; metilação do DNA; transcrição e informação acerca dos locais específicos de ligação de fatores, de forma a identificar um novo conjunto de elementos reguladores no DNA (57).

2. Análise do genoma humano

2.1. Ferramentas de análise bioinformática

A evolução tecnológica da sequenciação de genomas por NGS permitiu uma diminuição do custo da sequenciação mais do que seria expectável de acordo com a lei de Moore (aplicada inicialmente à evolução de *hardwares* computacionais e expandida a tecnologias como a de sequenciação) (58). Tal facto tem permitido a obtenção de um número crescente de genomas individuais. Com efeito, a fácil acessibilidade a uma grande quantidade de dados de sequenciação, contrasta com o desafio que recai na gestão e análise computacional desses dados (59). O alto rendimento das tecnologias de NGS faculta um considerável conjunto de dados genómicos, que carecem da utilização de ferramentas bioinformáticas para que possam ser dotados de significado. A bioinformática é assim a disciplina que veicula o entendimento do genoma sequenciado (58).

A análise de uma sequência genómica inicia-se pela criação de uma *pipeline* analítica, de forma a combinar métodos de análise para obter resultados biologicamente relevantes. Várias ferramentas bioinformáticas estão disponíveis para realizar cada uma das etapas de análise, sendo assim necessário estabelecer *a priori* a informação do genoma que se pretende obter, para uma correta escolha das ferramentas a utilizar. Depois de estabelecida uma *pipeline*, esta apresenta uma ordem pré-definida de etapas de análise e algoritmos internos que não devem ser modificados ou substituídos facilmente (Figura 1). O conjunto de etapas que constitui uma *pipeline* inicia-se pela garantia de qualidade das *reads* obtidas na sequenciação e a montagem do genoma através do alinhamento das *reads* contra um genoma de referência. Segue-se a identificação e a anotação de variantes para inferir a relevância biológica destas. Poderá realizar-se a prioritização e filtração das variantes identificadas para estudos funcionais futuros (60).

2.1.1. Montagem do genoma

A análise efetuada a jusante da sequenciação inicia-se pela avaliação da qualidade do produto dessa sequenciação. Algumas das *reads* obtidas diretamente do sequenciador apresentam uma qualidade relativamente baixa, e portanto, são sujeitas a remoção, corte ou correção aquelas que não cumpram as normas de qualidade definidas. Erros de leitura de bases, *reads* de má qualidade ou contaminação com adaptadores, são erros comuns no

produto da sequenciação, sendo que cada plataforma de sequenciação apresenta diferente distribuição e tipo de erros, dependendo da sua bioquímica de síntese (61). Atualmente, os sequenciadores não só produzem a informação da sequência de bases, como também produzem uma estimativa da probabilidade de erro para cada base identificada. Esta probabilidade é definida pelo "*Phred Quality Score".* A qualidade na escala *Phred* é definida em relação à probabilidade de erro P_e, por: Q Phred = - $10\log_{10} P_e$. O valor obtido permite inferir não só a probabilidade de erro como a exatidão de cada base (Tabela 1) (8, 62).

Tabela 1: Valores de qualidade *Phred.*

Q	P_e	Exatidão da base
10	1 em 10	90%
20	1 em 100	99%
30	1 em 1.000	99,9%
40	1 em 10.000	99,99%
50	1 em 100.000	99,999%

Após o processamento das *reads* para atender a um determinado padrão de qualidade, numa abordagem de re-sequenciação, estas são alinhadas contra um genoma de referência. Atualmente, o genoma de referência pode ser obtido através de duas fontes principais: *University of California Santa Cruz* (UCSC) (http://genome-euro.ucsc.edu) ou *Genome Reference Consortium* (GRC) do *National Center for Biotechnology Informatio*n (NCBI) (http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc). A UCSC apresenta a versão hg19 enquanto o GRC fornece a versão GRCh37. Ambos utilizam a mesma sequência de DNA consenso mas apresentam anotações diferentes. Inicialmente, a sequência de nucleótidos apresenta-se num ficheiro com formato FASTA, que consiste na sequência de nucleótidos sob a forma mais universal de representar sequências biológicas. É neste formato que é sempre apresentado o genoma de referência, contra o qual é feito o alinhamento. Com o aumento da capacidade dos sequenciadores os ficheiros neste formato passaram a ficar muito pesados, e uma forma de reduzir o volume de dados foi a criação do formato FASTAQ. Neste formato as bases e os valores de qualidade são apresentados no mesmo ficheiro (63).

Ao realizar o alinhamento das *reads* obtidas contra uma das referências referidas é importante denotar três questões: Ao mapear *reads* curtas num genoma de referência poderá existir um problema de ambiguidade, que pode em alguns casos ser ultrapassado com a utilização de *paired-end reads*; *reads* que só podem ser mapeadas com muitos *mismatches* não devem ser consideradas, o que faz com que variantes que são apenas suportadas por tais *reads* sejam descartadas; e tecnologias de segunda geração, que incorporam a PCR na preparação das bibliotecas, levam a que várias *reads* originárias numa única sequência molde possam ser sequenciadas, interferindo com a variável estatística, devendo-se assim remover duplicados de PCR após o alinhamento do genoma completo (60).

Após o alinhamento a informação é organizada em ficheiros com o formato *Sequence Alignment Map* (SAM), os quais armazenam a informação do alinhamento das *reads* com a referência. O formato *Binary Alignment Map* (BAM) é a versão comprimida em formato binário do ficheiro SAM (63). Vários programas de alinhamento podem ser utilizados para processar eficientemente milhões de *reads* curtas (ex. SOAP3-dp; Bowtie/Bowtie2; BWA) (20).

2.1.2. Anotação do genoma

A identificação de variantes é a etapa que sucede ao alinhamento das *reads* e depende da referência à qual essas *reads* estão alinhadas. Nesta etapa são identificados os locais, tipos e conteúdo das variantes específicas de um determinado genoma individual (63). As ferramentas bioinformáticas, para a realização da identificação, podem ser agrupadas em três categorias, de acordo com os tipos de variantes que detetam, nomeadamente SNPs e pequenos INDELs, SVs e CNVs. A identificação de variantes por uma abordagem de re-sequenciação é mais usada na procura de SNPs e INDELs, uma vez que são variantes que não requerem a eliminação de *reads* aquando do alinhamento. SVs e CNVs são significativamente mais estudadas em abordagens de sequenciação *de novo* (64). As dificuldades que o processo de identificação de variantes enfrenta prendem-se com a análise de regiões genómicas repetitivas; a identificação da fase do locus, ou seja, a origem materna ou paterna (importante em estudos de correlação genótipo-fenótipo); e a avaliação da cobertura da sequência, uma vez que se uma variação não for suportada por um

conjunto significativo de *reads* deverá ser descartada, levando a que alguns verdadeiros positivos possam ser excluídos (importante utilizar abordagens heurísticas para separar falsos positivos dos verdadeiros). As variantes são registadas num ficheiro de formato *Variant Call Format* (VCF), um formato criado pelo projeto dos 1000 Genomas, para guardar informações relativas às variantes detetadas de forma compactada e de fácil acesso (60).

Após a identificação da ocorrência de variantes, a anotação permite estabelecer quais as alterações que efetivamente provocam alterações funcionais e prever qual o seu valor no fenótipo. Com recurso a ferramentas de anotação é possível comparar as variantes com bases de dados existentes; prever *in silico* os efeitos das alterações; obter informação evolutiva (zonas conservadas); frequências alélicas; alterações na sequência de proteínas e a posição cromossómica relativamente a locais com interesse funcional (64). Um dos programas amplamente utilizado na anotação é o ANNOVAR (65). O ANNOVAR é uma ferramenta bioinformática para anotar SNPs e INDELs; determinar a sua consequência funcional, apresentando *scores* de importância funcional; encontrar variantes em regiões conservadas; e identificar variantes descritas no projeto dos 1000 Genomas ou na *The Database for Single Nucleotide Polymorphisms* (dbSNP), a base de dados de pequenas variantes genéticas do NCBI (53, 66). A anotação pode ser relativa a genes ou a outras zonas funcionais do genoma. Este tipo de anotação é importante para dados da sequenciação do genoma completo, uma vez que a maioria das variantes ocorre fora de regiões de codificação de proteína, não podendo os seus efeitos funcionais ser avaliados se a anotação fosse baseada exclusivamente em genes (65). O ANNOVAR utiliza ainda bases de dados de anotação do UCSC, *Ensembl*, ou quaisquer dados de anotação que sejam convertidos no formato *Generic Feature Format* 3 (GFF3), o formato de texto que representa anotações utilizado pelo ANNOVAR e no qual todos os ficheiros têm que ser convertidos (63). O ANNOVAR pode ainda avaliar e filtrar subconjuntos de variantes que não foram ainda descritas em bases de dados e são desconhecidas. Acresce a esta vantagem, muitas outras que distinguem este programa dos restantes, nomeadamente a possibilidade de aplicação a produtos de diversas plataformas de sequenciação e o facto de conseguir anotar as variantes presentes noutras regiões genómicas que não apenas génicas, como é exemplo regiões altamente conservadas, locais de ligação de fatores de transcrição e locais alvo de microRNAs (miRNAs) (65). No final da anotação, bem como no término

de cada etapa de análise, é efetuada uma validação e visualização dos resultados gerados. A curação manual da anotação de variantes individuais é fundamental para a prioritização de elementos genómicos, quando se pretende seguir para estudos funcionais experimentais (60) (Figura 1).

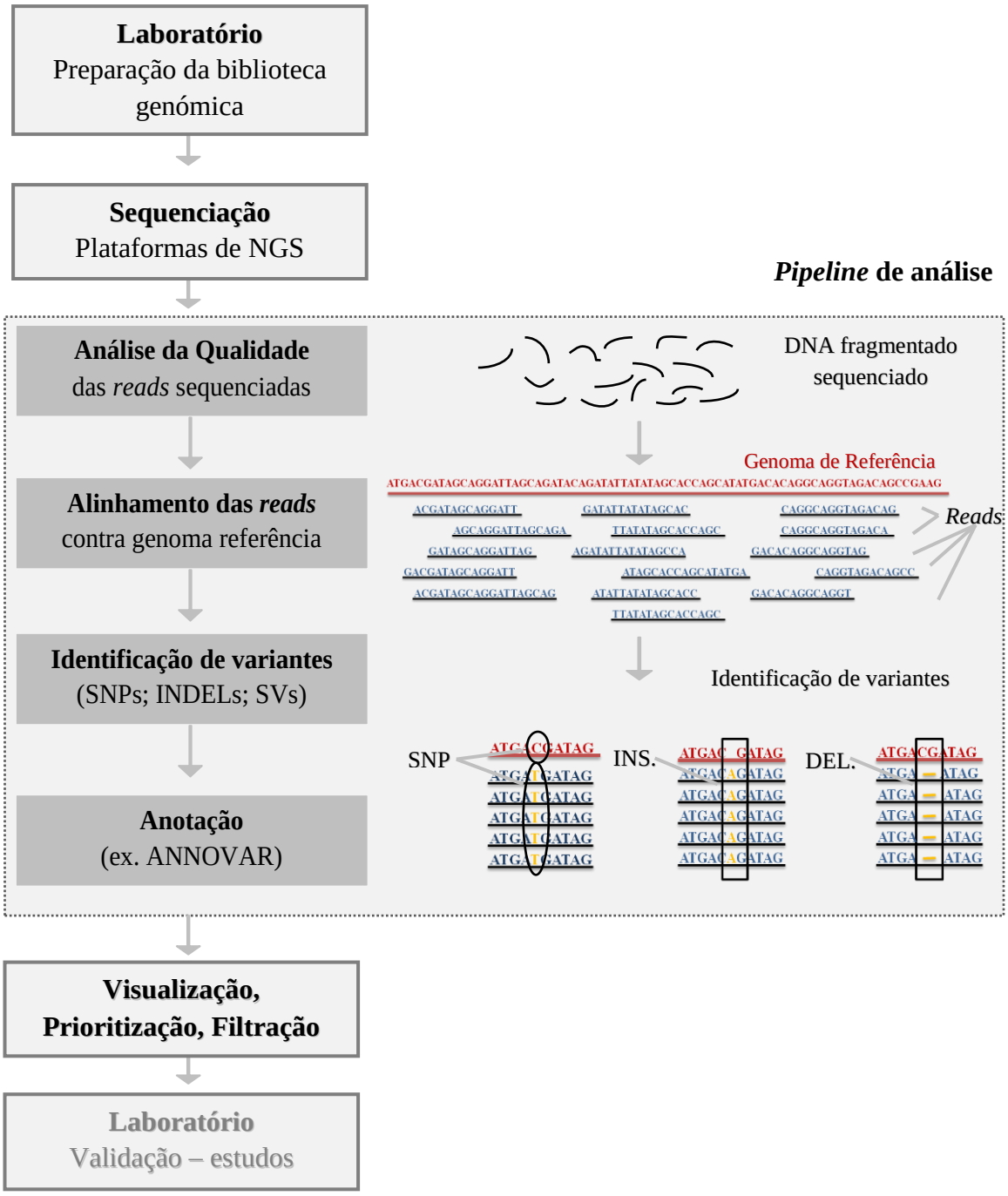


Figura 1: Possível fluxo de trabalho para a análise de genoma humano sequenciado por tecnologia de NGS.

2.2. Estrutura genómica de populações

A análise da variação genética humana tem permitido a compreensão da evolução do genoma humano, da história das populações, a caracterização da ancestralidade individual, a predisposição genética para a ocorrência de doença e o entendimento das vias biomoleculares desencadeadas por mutações, a principal fonte de variação genética individual (67).

Para compreender a variação individual de um genoma é necessário compreender a variação genética que ocorre no interior da população em que determinado indivíduo está inserido, bem como estabelecer um entendimento em relação à variação dessa população com as restantes populações. A genómica populacional, surge então como o estudo dos processos que afetam a evolução, distribuição e correlação das variantes genómicas dentro e entre populações (44). Na genómica populacional, uma questão de grande relevância no estudo da estrutura de populações é saber quanto diferem as populações geneticamente. A análise da variabilidade populacional, por meio de frequências alélicas das variantes genéticas, representa a melhor forma de responder à questão (68). A partir do valor de frequência alélica para determinados SNPs, é possível avaliar de que forma varia essa frequência entre populações, e para tal calcula-se o índice de fixação de *Wright* (F_{ST}) (Figura 2) (69). Este cálculo estatístico resulta da diferença entre a heterozigotia total (H_T) do conjunto de todas as populações em análise, e a média da heterozigotia ($\overline{H_S}$) de cada uma das populações. A heterozigotia observada (para um determinado locus) define-se como o número de indivíduos heterozigóticos observados, relativamente ao número total de indivíduos da população. O valor de F_{ST} indica claramente onde existe maior e menor variação, se entre as populações analisadas, ou se dentro de cada população. Quando este valor é igual a zero ($F_{ST} = 0$), a totalidade da variação existe dentro das populações, não existindo diferenças entre elas, ao passo que quando este valor é igual a um ($F_{ST} = 1$), a totalidade da variação existe entre populações, não existindo variabilidade no interior destas (69).

$$F_{ST} = \frac{H_T - \overline{H_S}}{H_T}$$

Figura 2: Cálculo do F_{ST} para a medida da diferenciação populacional relativamente à estrutura genética.

Assim, o F_{ST} poderá também ser considerado a medida da extensão em que as diferenças populacionais contribuem para a diversidade genética global. Jorde e colaboradores publicaram o resultado do cálculo estatístico para determinar a variação genética dentro e entre os continentes africano, europeu e asiático (70). Estes determinaram que cerca de 85% da variação se verificava dentro de cada continente, existindo aproximadamente 15% de variação genética entre os três continentes. É possível concluir que existe um elevado grau de variabilidade entre os indivíduos de uma população, o que diminui a diversidade entre populações e limita a capacidade de associar determinadas variantes a uma respetiva população. Analisando em detalhe esta conclusão é possível inferir que foram analisados SNPs com MAF>5%, portanto variantes comuns. Sendo uma variante relativamente comum, esta está presente no genoma humano há um período de tempo suficientemente grande para aumentar a sua frequência de ocorrência. Por outro lado, se for realizada uma análise a SNPs de baixa frequência ou raros, com MAF<5%, as percentagens invertem-se, existindo mais variabilidade entre os diferentes continentes. Uma variação rara é mais recente e como tal não teve tempo suficiente para a sua expansão. Assim, estas variantes são susceptíveis de serem específicas da população (68). Num estudo comparativo de estratégias para a caracterização da variação do genoma nas populações africana, asiática e europeia, foi possível denotar que para SNPs comuns (previamente identificados no dbSNP) a média da diferença da frequência alélica entre populações é cerca de 15% (Figura 3a). Já em SNPs raros, identificados *de novo* por sequenciação, apresentam aproximadamente 80% de diferenciação entre as três populações (Figura 3b) (49).

Atendendo ao facto das variantes raras estarem intimamente relacionadas com fatores causais de inúmeras doenças, facilmente se depreende que diferenças substanciais na frequência alélica destas variantes em diferentes populações, poderão refletir diferentes riscos para as mesmas (71). Estudos populacionais, como o de Jorde e colaboradores, realizam uma genotipagem com painéis de SNPs de alta frequência (70). Por contraste, para uma análise não só de variantes comuns como também das variantes raras presentes nas populações, é essencial avaliar a sequência completa do genoma, o que tem sido facultado pela tecnologia NGS (71). A partir da variabilidade genética entre populações é possível estabelecer uma métrica para a distância genética entre essas populações.

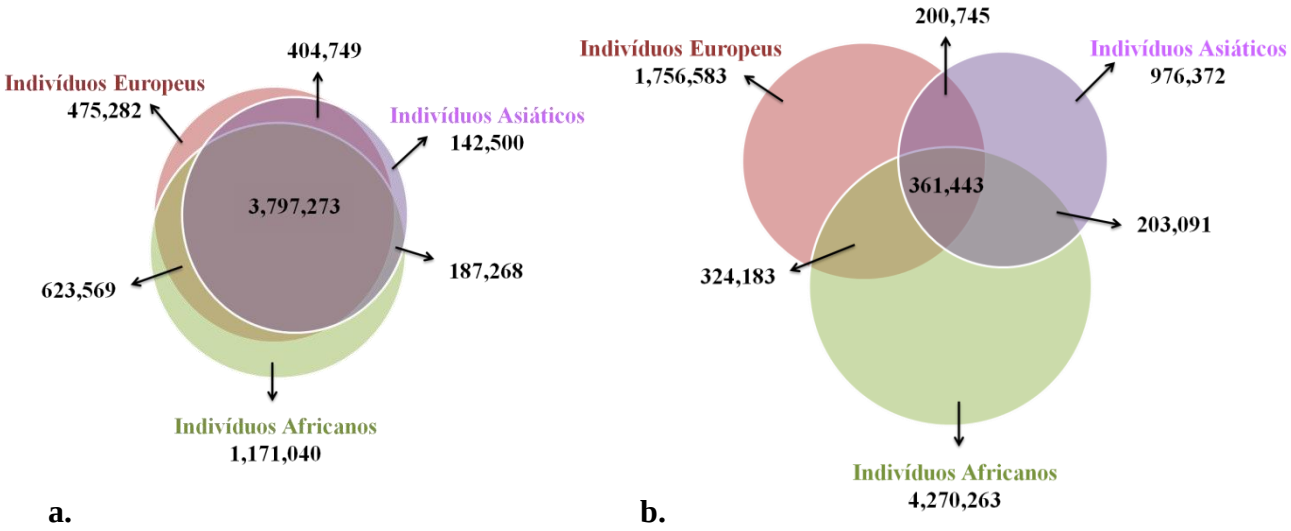


Figura 3: Diagrama de Venn que mostra o conjunto de SNPs partilhados pelos indivíduos do continente africano, asiático e europeu; **a.** Conjunto de SNPs comuns, identificados no dbSNP; **b.** Conjunto de SNPs raros, identificados *de novo* por sequenciação (adaptado de [48]).

A distância genética estatística resulta simplesmente da diferença nas frequências de alelos entre duas populações (72). Pela análise da distância genética entre diferentes populações é possível verificar um efeito de similaridade genética relativamente à localização geográfica em termos continentais (73). Este facto denota a falácia do conceito de uma população ser acasalada aleatoriamente. Com efeito, os indivíduos são mais propensos a ter descendentes com alguns indivíduos em detrimento de outros. A distância geográfica e as fronteiras nacionais são barreiras importantes para o fluxo de genes, tal como a própria inserção em grupos socioeconómicos e culturais também o delimitam. A correlação entre a distância genética e a localização geográfica pode resultar na similaridade de localização genética verificada nas populações de cada continente. Uma outra conclusão retirada do estudo de distâncias genéticas é o facto da população africana apresentar a distância mais curta do ancestral populacional, como tal acredita-se que representa a população parental, ou seja a origem da população do resto do mundo. Da análise de distâncias é possível não só determinar a origem da população humana, como caracterizar todo o movimento migratório e expansão para os restantes continentes (Figura 4). Mais, é possível caracterizar uma distribuição biogeográfica das populações humanas, e não obstante da variabilidade genética existente em cada continente, a combinação de informação de um número considerável de loci poderá determinar a ancestralidade individual (73).

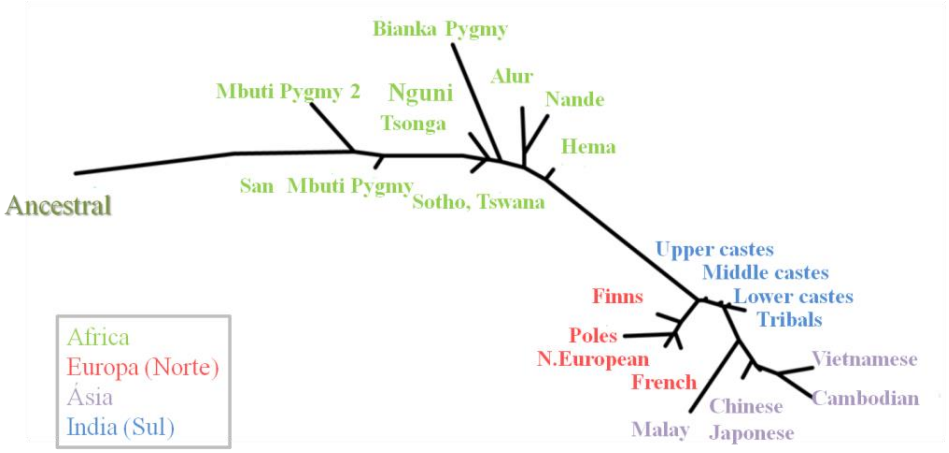


Figura 4: Distância genética calculada a partir de 100 polimorfismos de inserção *Alu*, em populações do continente africano, europeu, asiático (adaptado de (73)).

2.3. Ancestralidade

Com o advento da era genómica foi possível separar o conceito de ancestralidade biogeográfica do conceito sociocultural de Etnia (Raça) (74). No passado, quando a acessibilidade a estudos genómicos era limitada, a caracterização da ancestralidade individual era obtida através da identidade racial auto declarada do próprio indivíduo, causando inevitavelmente incertezas na estrutura genómica da população subjacente (75). Geneticamente, não existe o conceito de raças humanas no sentido de agrupamentos inequívocos discretos, sendo essa uma construção social. O facto do valor médio F_{ST} ser de aproximadamente 15% entre continentes, vem corroborar a inadequação da utilização deste conceito a nível genético (70). Decorrente da descendência de uma população parental e mobilidade populacional, patentes na evolução da história populacional, a ancestralidade genética surge a partir da distribuição biogeográfica das populações humanas. O estudo da ancestralidade genética permite não só caracterizar a estrutura populacional em que o indivíduo se insere, definindo o número de subpopulações presentes e atribuindo o indivíduo a uma determinada subpopulação; como também determina o número de populações ancestrais da população e atribui as proporções de cada ancestral populacional a um indivíduo, identificando neste a ancestralidade genética de segmentos cromossomais distintos (37, 41, 42).

Com o conhecimento prévio das frequências alélicas específicas da população, podem ser utilizados painéis de SNPs referidos como marcadores informativos de ancestralidade (AIMs), para estimar a ancestralidade genética (76). Os AIMs são marcadores cuja frequência é significativamente diferente, o que torna possível distinguir duas ou mais populações. Alguns conjuntos de AIMs podem ser selecionados a partir de GWAS ou utilizando uma abordagem de análise dos componentes principais (PCA). Os painéis destes marcadores variam em tamanho, mediante a sua finalidade. Painéis relativamente pequenos, com um número de dezenas a centenas de SNPs, podem ser utilizados quando é pretendido definir a ancestralidade genética continental, ao passo que centenas a milhares de SNPs são necessários para uma estimativa de subpopulações dentro dos continentes (76). A ancestralidade genética pode ainda ser dividida em estimativas locais ou globais. As estimativas locais procuram identificar a origem ancestral dos segmentos cromossomais dentro de um genoma individual. Estimativas globais, por seu lado, procuram estabelecer proporções médias de ancestralidade ao longo de todo o genoma individual, de forma a atribuir as proporções de cada ancestral a um determinado indivíduo. Em 2014, Wang e colaboradores apresentaram o *Locating Ancestry from Sequence Reads* (LASER), uma ferramenta bioinformática que permite caracterizar um genoma individual quanto à sua ancestralidade genética global (77). Utilizando uma abordagem de PCA, o LASER permite obter uma estimativa de ancestralidade continental, bem como de subpopulações que existem dentro dos continentes, nomeadamente na Europa, o continente com a variação genética mais homogénea (78).

2.4. Análise de risco

A variação genética que pode influenciar a saúde dos indivíduos tem sido um foco central de investigação desde a introdução do conceito de "individualidade química" (46). Atualmente, com a disponibilidade da sequência do genoma humano e com a determinação da extensão da variação do genoma, dentro e entre populações, a consciência de que cada indivíduo tem a sua própria variação genética única, fornece a base conceptual da medicina personalizada. A perspetiva de uma medicina individualizada prende-se com a identificação genotípica de fatores de risco para a ocorrência futura de determinadas doenças e a modulação da decisão terapêutica com base na genética individual. Assim, tecnologias de NGS fornecem informações que, em combinação com dados clínicos e

análise de risco ambiental, contribuem para a orientação da tomada de decisão clínica. A análise de risco pode focar-se nas variantes associadas a genes para doenças mendelianas; nas mutações identificadas *de novo*; nas variantes conhecidas que modulam a resposta à terapia farmacológica e/ou nas variantes anteriormente associadas com doenças complexas (44). Para determinar uma estimativa do risco são aplicados algoritmos de previsão para ponderar a probabilidade de patogenicidade de determinada variante com base na frequência alélica, conservação e alteração do domínio proteico (71). Para tal, recorre-se a bases de dados de mutações específicas de doenças como o *Online Mendelian Inheritance in Man* (OMIM) ou o *Human Genome Mutation Database* (HGMD), e a bases de dados de farmacogenómica, como a *Pharmacogenomics Knowledge Base* (PharmaGKB) (79, 80). O estudo de variantes implicadas na modificação da metabolização e resposta a fármacos teve um grande impulso com a identificação de variantes comuns por GWAS. No entanto, embora a contribuição genética para a resposta farmacológica resulte também da contribuição de variantes comuns, as variantes mais relevantes demonstram uma baixa frequência alélica, o que é possível detetar pela análise do genoma sequenciado. Em 2010, Ashley e colaboradores sequenciaram o genoma de um indivíduo com história familiar de doença vascular e morte súbita (81). O estudo identificou variantes genéticas associadas ao aumento do risco genético para o enfarte do miocárdio, diabetes tipo 2, cancro e morte súbita cardíaca. Concomitantemente, o mesmo estudo identificou mutações em genes relacionados com a resistência a clopidogrel (anticoagulante plaquetar, utilizado na prevenção do enfarte do miocárdio), em genes associados a uma resposta positiva a hipolipemiantes (agentes hipolipidémicos, utilizados no tratamento de dislipidemias e doenças vasculares) e genes relacionados com a varfarina (anticoagulante, especialmente relevante em farmacogenómica devido à sua estreita janela terapêutica).

2.5. Doenças mendelianas

Uma das principais aplicações do estudo da sequência completa do genoma humano é a procura de variantes responsáveis por doenças causadas por um único gene (mendelianas), ou características monogénicas. A análise de variantes genéticas responsáveis por uma doença monogénica é necessária, não só para diagnosticar definitivamente a doença, como para fornecer informação prognóstica e aconselhamento genético familiar, estabelecendo o risco de ocorrência para outros indivíduos (60). Os

estudos de doenças monogénicas, tanto autossómicas recessivas como a fibrose quística, ou dominantes como a neurofibromatose tipo 1, revelam mutações nos genes únicos para a ocorrência e transmissão da doença. Contudo, muitos dos fenótipos de doenças mendelianas são geneticamente heterogéneos. Doenças mendelianas como a surdez ou a retinite pigmentosa têm mais de 100 genes com mutações causais identificadas. Acresce ainda o facto de mutações específicas poderem conferir o fenótipo em homozigotia, em heterozigotia ou mesmo em heterozigotia composta (82). Assim, mesmo quando existem padrões simples de hereditariedade em doenças com um curso patológico bem caracterizado, os eventos mutacionais subjacentes precisam ser resolvidos para se obter um diagnóstico molecular preciso. Com efeito, a tecnologia NGS veio permitir o estudo de muitas alterações monogénicas já descritas. Atualmente, o OMIM é o catálogo onde o resultado de estudos de doenças mendelianas é descrito, sendo que deste já constam mais de 3.000 doenças em que foi avaliada a base biomolecular. Mais de 3.500 doenças, porém, mantêm a causa genética desconhecida e carecem de identificação (79).

Em 2010, Lupski e colaboradores publicaram a primeira demonstração da utilização da sequenciação do genoma completo, numa abordagem clinicamente relevante (83). O genoma completo do próprio James Lupski, um dos autores do estudo, foi sequenciado por NGS, com a finalidade de estudar a etiologia genética da doença mendeliana de Charcot-Marie-Tooth, que o próprio detém. A doença Charcot-Marie-Tooth é uma doença desmielinizante que apresenta heterogeneidade genética, uma vez que já foram identificados mais de 30 genes com mutações causais, quer de herança autossómica dominante, recessiva ou ligada ao cromossoma X. Porém, o genoma sequenciado não apresenta nenhuma das alterações génicas causais identificadas. Deste modo, a análise centrou-se no conjunto de variantes identificadas e ainda não descritas. O número elevado de alterações identificadas tornou necessário focar o estudo apenas nas variantes que ocorrem nos genes previamente ligados à doença. Dada esta filtração, foi possível identificar duas mutações, uma em cada alelo do gene *SH3TC2*, que se expressa nas células de Schwann que envolvem a bainha de mielina em torno dos nervos. A vantagem proporcionada pela análise de todo o genoma individual sequenciado permitiu caracterizar a alteração molecular conducente à doença, e caracterizar toda a família para a ocorrência ou não da doença.

A complexidade do estudo de doenças monogénicas, geneticamente heterogéneas, aumenta quando perante dois fenótipos, um na presença e outro na ausência de doença, os genomas sequenciados não apresentam alterações na sua estrutura génica. Kingsmore e colaboradores sequenciaram o genoma de dois gémeos monozigóticos, em que um deles tinha a doença autoimune de esclerose múltipla e o outro não. Os genomas monozigóticos discordantes para a esclerose não apresentavam quaisquer diferenças genéticas (84). Deste modo, a análise do genoma humano sequenciado tem também a capacidade de eliminar possibilidades diagnósticas, veiculando outras análises funcionalmente relevantes, como o estudo do epigenoma.

2.6. Doenças complexas

Um crescente interesse na análise da sequenciação do genoma completo tem sido a descoberta da base molecular de doenças complexas, altamente prevalentes, como a Diabetes Mellitus, doenças cardiovasculares, obesidade, autismo ou cancro, doenças que resultam da combinação de várias alterações genéticas e fatores ambientais (85). Dois modelos distintos direcionam a análise genómica de variantes nas doenças complexas (86). O modelo, inicialmente seguido, reside na hipótese "Variação Comum – Doença Comum", em que os fenótipos complexos são o resultado de efeitos cumulativos de um grande número de variantes comuns. De acordo com este modelo o conjunto de 10 a 15 milhões de SNPs, CNVs e outras variantes, que são comuns na população, são moduladas pela presença de fatores ambientais em termos de hereditariedade. A realização de GWAS encontrou um conjunto de associações entre variantes comuns e fenótipos complexos que corroboram a hipótese (87). Contudo, a perda de herdabilidade verificada pelas variantes encontradas, levou à necessidade de atender a um modelo alternativo com vista a detetar variantes causais das condições avaliadas (52). Este modelo alternativo focou-se na hipótese "Variação Rara – Doença Comum", e argumenta que a suscetibilidade genética para as doenças complexas é devida ao risco acumulado conferido por múltiplas variantes raras no genoma de um indivíduo, portanto variantes que não são passíveis de serem detetadas por estudos de variantes comuns. O estudo das frequências mais baixas das variantes raras é possível pela análise do genoma sequenciado (85). Tal abordagem já foi utilizada em situações como na doença de Parkinson (88), doença cardiovascular (89) ou diabetes (90). A utilização de NGS tem também levado à identificação de mutações que

conduzem a vários tipos específicos de cancro, que são descritas como estruturais ou não codificantes (91). A sequenciação abriu caminho para a compreensão de genes alterados em células cancerígenas, de vias alteradas e como esses dados contribuem para os modelos de estudo com finalidade de conhecimento da bioquímica do cancro (revisto em (92)). Recentemente, um estudo respeitante à vacinação contra tumores, mostrou a capacidade de induzir imunogenicidade através de mutações somáticas detetadas por NGS, em linfócitos T (93). Dado o grande número de fatores etiológicos genéticos e não genéticos das doenças complexas, a abordagem de estudo irá sempre necessitar de integração de dados adicionais, nomeadamente da transcriptómica, proteómica e metabolómica (44).

3. RNAs não codificantes

3.1. Novos elementos funcionais

Com o início do desenvolvimento das tecnologias de sequenciação aliado à vontade expressa pela comunidade científica em analisar a informação genética individual, a sequenciação de exomas (≈ 2% do genoma), ao invés de genomas completos, mostrou ser uma alternativa à limitação dos elevados custos da tecnologia de então. Assim, foi viabilizado o estudo da matéria codificante do genoma, a qual se julgava ser a única região genómica dotada de importância funcional. Com efeito, durante décadas, a principal função celular reconhecida ao ácido ribonucleico (RNA) foi a de molécula intermediária na transferência da informação genética do DNA para as proteínas. Não obstante do facto de a transcrição ser um processo abrangente à maioria do genoma, no término do PGH o genoma sequenciado revelou apenas ≈21.000 genes codificantes de proteínas, pouco mais que do que apresenta o genoma de *Caenorhabditis elegans* (*C. elegans*) (20.060) ou da *Drosophila melanogaster* (14.039) (13). O genoma humano, cerca de 30 vezes maior que o genoma destes dois organismos, revelou uma maior complexidade funcional quando foi analisado em detalhe o material transcrito (revisto em (94)). O advento da tecnologia NGS associado ao avanço necessário na área da bioinformática, permitiu ao projeto ENCODE caracterizar os transcritos do genoma humano e encontrar novos elementos funcionais que revolucionaram a noção de função celular até então atribuída ao RNA (56). O genoma humano codifica dezenas de milhares de RNAs que não codificam proteína, o que necessariamente sugere que esta fração não codificante do genoma desempenha uma função chave, que confere complexidade à fisiologia humana. Notavelmente, a descoberta

em 1993 da função de silenciamento da expressão génica exercida por pequenas moléculas de RNA não codificante, abriu caminho à caracterização funcional destas moléculas (95). Já em 2011, foi possível revelar regiões do genoma que codificam longos RNAs não codificantes que também denotam importantes funções biológicas (96). Assim, é possível distinguir pequenos RNAs não codificantes, moléculas de ≈20 a 30 nucleótidos de comprimento, de longos RNAs não codificantes, moléculas de comprimento superior a 200 nucleótidos (97, 98).

A importância funcional evidenciada pelas moléculas de RNA não codificante realça a mais valia patente no estudo do genoma completo, ao invés do estudo exclusivo do exoma humano, para a caraterização funcional. Na verdade, a diminuição dos custos de sequenciação associada à evolução das técnicas de NGS, tem tornado este estudo possível. A descoberta de sequências de genes de RNA não codificante torna-se difícil de efetuar apenas por previsão de métodos computacionais, dada a diversidade evolutiva destas moléculas, portanto a incorporação da tecnologia de sequenciação no seu estudo tornou-se relevante. A caracterização destas moléculas melhora a anotação dos genomas de tal modo que o impacto das mutações será amplamente interpretável em todo o genoma (18).

3.2. MicroRNAs

Os miRNAs são uma classe de sequências de ≈18 a 25 nucleótidos de RNA não codificante, os quais regulam pós-transcripcionalmente a expressão génica, pela inibição ou degradação de mRNAs alvo (99). Estas moléculas oligoribonucleotídicas passaram despercebidas até aos anos 90, porém desde a sua descoberta em 1993, o seu estudo abriu uma nova era de entendimento na regulação celular, acrescentando um novo mecanismo ao dogma central da biologia molecular (100). Lee e colaboradores descreveram pela primeira vez um RNA de 22 nucleótidos, regulador da expressão génica em *C. elegans*, codificado pelo gene lin-4, o qual deu nome ao primeiro miRNA descrito (95). Apenas no ano de 2000, Reinhart e colaboradores descobriram o segundo miRNA, let-7, em *C. elegans*, sendo mais tarde descrito numa variedade de células animais, sugerindo a distribuição ubíqua dos miRNAs (101). Em 2001 foi formalmente introduzido o termo miRNA e desde então têm sido identificados em plantas, animais e vírus, como reguladores negativos dos seus genes alvo através do emparelhamento com a região 3'UTR do mRNA, destabilizando-o e inibindo a tradução deste (99).

Os miRNAs têm demonstrado ser a classe de moléculas reguladoras mais abundantes no ser humano, e até à data cerca de 1.500 miRNAs foram identificados. Estima-se que 30% de todos os genes estão sujeitos à sua regulação (99). A complexidade da regulação génica via miRNAs surge logo na regulação da expressão dos próprios miRNAs. Fatores de transcrição chave como o fator indutor de hipóxia (HIF), o oncogene *c-myc*, a proteína supressora tumoral *p*53 e o fator nuclear Kappa B (NF-kB), já foram descritos como reguladores de *clusters* de miRNAs (102, 103). No entanto, a análise dos fatores de transcrição apresenta um enorme desafio pelo facto do local de ligação destes às regiões promotoras dos miRNAs se encontrar a uma distância que pode ir de poucos Kb até mais de 50 Kb a montante dos genes de miRNAs (104). Mais, o facto de um dado fator de transcrição poder regular um *cluster* de miRNAs, os quais por sua vez podem modular outros fatores de transcrição pelos seus genes alvo, conduz à formação de circuitos genéticos, de grande complexidade reguladora na síntese de proteínas. Dado o enorme envolvimento em vias genéticas de produção proteica, os miRNAs regulam inúmeros processos, nomeadamente diferenciação celular, apoptose, resposta imune, processos neoplásicos e diversos processos metabólicos. A necessidade destas moléculas para o normal funcionamento fisiológico, bem como os processos fisiopatológicos que resultam da sua desregulação, motivam a identificação e caracterização de miRNAs no genoma humano (105).

Os métodos convencionais como o *Northern Blotting* e a reação em cadeia da polimerase quantitativa após transcrição reversa (qRT-PCR) têm sido utilizados no estudo de miRNAs, contudo estes denotam limitações significativas em resultado do curto comprimento nucleotídico destas moléculas. Assim, tecnologias de alto rendimento como NGS têm facilitado a sequenciação de tais moléculas, à escala genómica (105). O grande volume de dados produzido pela sequenciação necessita então de abordagens *in silico* para ajudar a determinar os loci genómicos, fatores de transcrição de regiões promotoras, funções e alvos de miRNAs. Deste modo, nos últimos anos, o estudo destas moléculas tem sido alvo de especial consideração, tendo surgido em Setembro de 2009 a miRBase, uma base de dados de miRNAs que já contém descritos cerca de 24.521 miRNAs de 206 espécies (106).

3.2.1. Biogénese de microRNAs

Os miRNAs têm origem no núcleo celular, onde os genes que os codificam são transcritos, gerando uma sequência primária, o pri-miRNA. Este percursor, com um comprimento entre 60 a 80 nucleótidos é, geralmente, transcrito por uma RNA polimerase II (excecionalmente, pela polimerase III na transcrição de miRNAs presentes em regiões repetitivas de elementos *Alu*), mantendo no transcrito estruturas de mRNA como a Cap 5' e a cauda 3' poli A. O pri-miRNA transcrito dobra sobre si próprio formando uma estrutura em *hairpin*. No núcleo esta estrutura é clivada por ação combinada de uma enzima RNase III Drosha e pelo seu fator de ligação *DiGeorge critical region 8* (DGCR8), convertendo o pri-miRNA em pré-miRNA (107). O pré-miRNA contém um comprimento ~55 a 70 nucleótidos, sendo reconhecido o local de corte da enzima Drosha pela Exportina-5, uma proteína que permite a translocação do pré-miRNA do núcleo para o citoplasma (108). Já no citoplasma, o pré-miRNA é novamente processado por uma RNase III, a enzima Dicer, que por intermédio de uma nova clivagem forma um miRNA de cadeia dupla, com um comprimento de ~18 a 25 nucleótidos (109). As duas cadeias do miRNA formado são separadas por uma RNA helicase, permanecendo a cadeia cuja extremidade 5' liga mais fracamente com a cadeia complementar como miRNA maduro, ao passo que a cadeia complementar torna-se na cadeia passageira miRNA*, que é rapidamente degradada. Sucessivamente, em conjunto com proteínas Argonaute (Ago, proteínas que contêm três domínios conservados PAZ, MID e PIWI), e GW182 (proteínas de 182 kDa que contêm elevado teor de glicina e triptofano – GW), o miRNA maduro forma um complexo ribonucleoproteíco de silenciamento, denominado miRISC. As proteínas Ago são a componente catalítica do miRISC, sendo que a sua atividade de endonuclease cliva o mRNA no complexo miRNA-mRNA, não clivando o mRNA quando este não está associado a um miRNA (110). Esta é a via canónica de formação de miRNAs, ativada maioritariamente para a expressão de miRNAs presentes em regiões intergénicas do genoma. Contudo, alguns miRNAs resultam de um processamento diferente. Cerca de um terço dos genes de miRNAs está localizado em intrões de genes codificantes de proteínas e o seu processamento ocorre via *splicing*, sendo esta a via miRtron. Esta via conduz à formação do pri-miRNA por meio de *splicing* das regiões intrónicas onde genes de miRNAs estão presentes, não requerendo a ação do complexo Drosha/DGCR8 na formação do pré-miRNA. O pri-miRNA que resulta do *splicing* não adquire

automaticamente a estrutura de *hairpin*, este é então processado por uma enzima de fragmentação que permite adquirir a conformação de pré-miRNA e ser transferido pela Exportina-5, para posterior processamento via Dicer e formação do miRISC, tal como sucede na via canónica (111) (Figura 5).

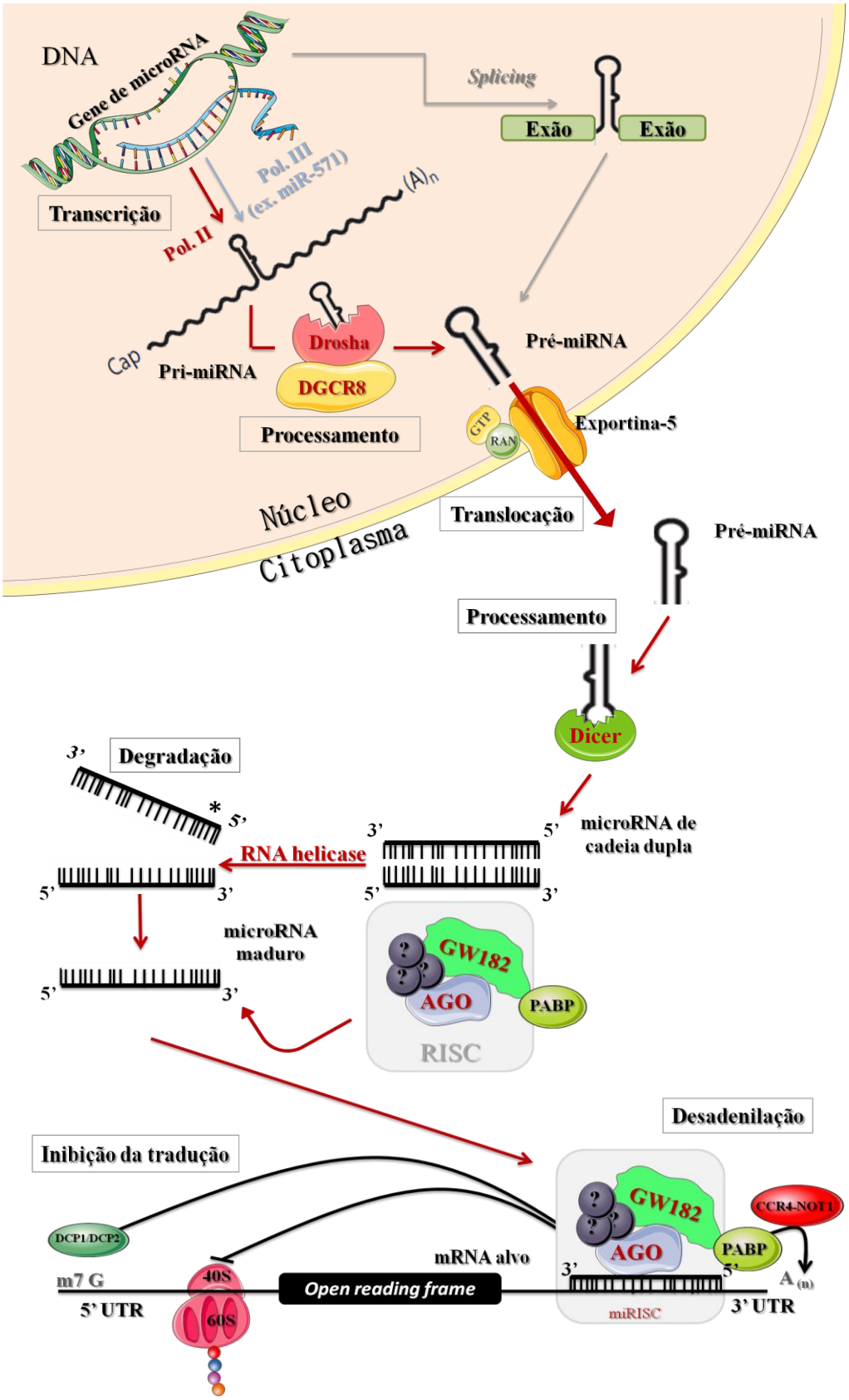


Figura 5: Processo de formação de miRNAs (↑Via Canónica; ↑Via miRtron; ↑Via RNA polimerase III). Ligação da molécula formada a um mRNA alvo de forma a regular negativamente a expressão deste.

3.2.2. Caracterização do complexo miRNA – mRNA

A ação primária dos miRNAs ocorre na forma de complexos ribonucleoproteícos miRISC e consiste na sua associação a um mRNA alvo, num processo conseguido através do emparelhamento de bases. A característica necessária para o reconhecimento de um mRNA alvo por parte de um complexo miRISC, é a complementaridade da região 3' UTR do mRNA com a região *seed* do miRNA, um segmento específico entre os nucleótidos 2 a 8 (110). O primeiro nucleótido da extremidade 5' do miRNA, na forma monofosforilada, liga-se aos aminoácidos conservados da proteína Ago (entre os domínios MID e PIWI) de forma direta ou por interação de um ião de magnésio. Assim, os nucleótidos seguintes, da posição 2 à 6 do miRNA, contactam com a Ago através da cadeia fosfatídica das riboses, sendo apresentados à superfície proteica numa conformação semi-helicoidal, com as bases disponíveis para a ligação de hidrogénio com o mRNA alvo. Estas propriedades são então responsáveis pelo facto do nucleótido na posição 1 não precisar de emparelhar com o alvo, e pelo facto da complementaridade na região *seed* ser crucial para que se forme o complexo miRNA-mRNA (112). Contrariamente ao que ocorre nos miRNAs das plantas, onde a complementaridade com o transcrito alvo é essencialmente exata, nos miRNAs humanos a complementaridade é incompleta e pode conter *mismatches* e protuberâncias, desde que estes não ocorram na região em que o miRNA está associado com as proteínas Ago. A complementaridade próxima da extremidade 3' do miRNA, nomeadamente entre os nucleótidos 13 a 16, contribui para a estabilidade do complexo miRNA-mRNA, ocorrendo mais *mismatches* na região central do miRNA, essencialmente nas posições 10 a 12. Efetivamente, dependendo da extensão da complementaridade, os miRNAs podem exercer um de dois efeitos no mRNA alvo: a clivagem e degradação do mRNA (maior grau de complementaridade) ou a inibição da tradução do mRNA (menor grau de complementaridade). Independentemente do mecanismo seguido, ambos são conducentes a uma diminuição da síntese proteica (113).

Esta supressão da tradução de mRNAs em sequências proteicas, motivada por miRNAs, pode ocorrer por intermédio de efeitos diretos ou indiretos. A inibição da tradução de forma direta ocorre quando o complexo miRISC inibe o reconhecimento da subunidade 40S ribossomal pela extremidade Cap 5' do mRNA, ou quando este inibe a junção da subunidade 60S, antagonizando a formação do complexo ribossomal 80S,

necessário à síntese proteica (114). Caso a tradução já tenha sido iniciada, o complexo miRISC inibe diretamente a elongação do ribossoma e promove a proteólise de polipéptidos já formados (115). A inibição indireta da tradução, exercida pelo complexo miRISC, ocorre pela destabilização e subsequente degradação ou compartimentalização de mRNAs alvo. Para que a destabilização do mRNA ocorra, o complexo miRISC interage com o complexo de desadenilação CCR4-NOT1, para remover a cauda poli A da extremidade 3' do mRNA, que por sua vez está ligada ao miRISC pela interação da sua proteína de ligação PABP e a proteína GW182 associada à Ago no miRISC. Depois da desadenilação, a estrutura Cap da extremidade 5' do mRNA é removida pelo complexo enzimático DCP1-DCP2. O que resta do mRNA alvo pode ser imediatamente degradado pela exonuclease XRN1, no sentido 5' – 3', ou pode ser acumulado em regiões citoplasmáticas denominadas de *Processing bodies* (*P bodies*), locais que contêm enzimas necessárias à degradação de mRNAs que aí são acumulados bem como ao *turnover* de miRISCs (110).

3.2.3. Relevância clínica: O premente desafio do estudo

Os miRNAs regulam mRNAs, os quais por sua vez codificam proteínas que efetivam funções celulares por diversas vias moleculares. É desta forma que estas moléculas reguladoras desempenham um papel vital na viabilidade celular, influenciando desde o processo de diferenciação celular, passando pela proliferação, até ao processo de morte celular. A primeira evidência que denota a relevância dos miRNAs a nível fisiológico, surgiu em 2003, quando Bernstein e colaboradores inibiram a formação global de miRNAs, pelo silenciamento da enzima Dicer, em células estaminais de ratinho, e verificaram a morte do embrião (116). Estudos posteriores confirmaram a necessidade de miRNAs para a sobrevivência e desenvolvimento em diversos estadios celulares. Certos miRNAs apresentam padrões de expressão característicos, nomeadamente específicos de determinados tecidos, como é exemplo o miR-21 no coração, o miR-1 no músculo e o miR-122 no fígado, ou específicos de determinado estadio de desenvolvimento, como o miR-290 que é altamente expresso em células estaminais ou o miR-143 elevado nos adipócitos durante a adipogénese. Outros, porém, apresentam uma expressão global, nomeadamente os que regulam a expressão de genes que codificam proteínas de distribuição ubíqua (revisto em (117)).

Tendo conhecimento dos padrões de expressão temporal, é possível caracterizar assinaturas de miRNAs de um *status* fisiológico específico, ou seja um perfil de miRNAs expressos em determinada condição e que contribui para a homeostasia celular. No entanto, tal como existe variação genética que pode alterar a expressão e/ou função de determinada proteína, também a presença de variação em genes que codificam miRNAs, nos locais alvo ou nos próprios fatores de transcrição de miRNAs (os quais, muitas vezes, são também eles regulados por outros miRNAs), pode resultar na sobre ou sub-expressão de determinado miRNA e consequentemente na alteração da expressão dos seus alvos. Assim, é possível obter assinaturas de miRNAs, não só de estados de homeostasia fisiológica como de doença, uma vez que a alteração da sua expressão pode conduzir ou coadjuvar a ocorrência de uma condição fisiopatológica. Com efeito, o conhecimento do perfil de um estado de homeostasia permite já antever o resultado da sua desregulação no funcionamento celular e sistémico, onde determinado miRNA está inserido, bem como a desregulação indireta que esses miRNAs provocam noutros sistemas regulados pelos seus alvos. Atendendo, por exemplo, aos miRNAs que regulam o metabolismo e a homeostasia energética, é possível verificar que estes são importantes, não só na diferenciação dos tecidos envolvidos na produção e armazenamento de energia (fígado, músculo esquelético e adipócitos), como também regulam a libertação de insulina, o metabolismo de aminoácidos e lípidos. Da caracterização prévia do perfil de expressão fisiológico neste sistema metabólico, foi possível prever e posteriormente confirmar experimentalmente, que a sobreexpressão do miR-375 leva à diminuição da secreção de insulina pelas células β pancreáticas, e a diminuição da expressão deste conduz ao aumento da expressão de insulina (118). Tal pode acontecer no metabolismo dos aminoácidos, onde o miR-29b tem como alvo o transcrito da desidrogenase dos cetoácidos de cadeia ramificada (BCKD), conduzindo a sua desregulação à alteração da expressão normal desta enzima (119). A BCKD catalisa o primeiro passo irreversível no catabolismo dos aminoácidos de cadeia ramificada (BCAAs), ou seja a leucina, isoleucina e valina, aminoácidos essenciais. Como a leucina consegue estimular a secreção de insulina, a regulação direta que o miR-29b efetua na BCKD faz com que este miRNA regule indiretamente o metabolismo da insulina, resultando da sua desregulação a alteração das duas vias metabólicas.

O principal alvo de investigação na caracterização de perfis de miRNAs com expressão alterada reside, porém, no processo neoplásico. Inicialmente Lu e colaboradores

detetaram a diminuição de miRNAs através da análise de centenas de amostras de tecido neoplásico, indicando que estas moléculas teriam um papel de supressor tumoral (120). Atualmente, os tecidos tumorais apresentam também padrões de sobreexpressão destas moléculas, motivando a regulação negativa de supressores tumorais e portanto exacerbando a ação de oncogenes, sendo por isso descritos como oncomiRs. O miR-21, a primeira molécula descrita como oncomiR, é um dos miRNAs sobreexpresso, quer em cancro da mama, pâncreas e pulmão, promovendo a mobilidade celular e invasão, uma vez que atua diretamente na proteína PTEN, um supressor tumoral que inibe a invasão celular bloqueando metaloproteinases da matriz (121). Estes estudos têm demonstrado o potencial da utilização da assinatura de miRNAs no diagnóstico do cancro, constituindo vários miRNAs como biomarcadores de diagnóstico e prognóstico tumoral (120).

Por outro lado, a alteração da expressão dos miRNAs não significa, exclusivamente, a ocorrência de doença, esta pode também ocorrer em resposta a um estímulo ou alteração da fisiologia celular, promovendo alteração na expressão das suas moléculas alvo de modo a que a alteração confira proteção celular. Tal facto é evidenciado na regulação de sistemas críticos para a homeostasia, como a regulação do *stress* oxidativo no sistema cardiovascular. Tem sido demonstrado que os miRNAs afetam a resposta redox em células endoteliais, através da regulação do fator de transcrição HBP1 e da consequente *p*47, vital para a ativação da NADPH oxidase (122). Cheng e colaboradores demonstraram que radicais livres induzem a expressão de miR-21, o qual leva à proteção de cardiomiócitos do *stress* oxidativo, pela inibição da expressão do gene PDCD4 (123).

A caracterização dos perfis de expressão, bem como o estudo de novos miRNAs e respetivos locais alvo, permite um entendimento cada vez maior da rede de interações genéticas em que estas moléculas estão envolvidas. O facto de abordagens de previsão de miRNAs *in silico* apresentarem uma precisão relativamente baixa, com um número de falsos positivos de, sensivelmente, metade dos miRNAs previstos por cada análise, faz com que a análise genómica possibilitada por tecnologias de sequenciação, seja premente para desvendar novas vias de regulação génica, prever e diagnosticar doenças, bem como alcançar terapêuticas eficazes baseadas na regulação destas moléculas (124).

Objetivos

O estudo centra-se na análise de sequências completas do genoma humano de quatro indivíduos de nacionalidade portuguesa, obtidas por tecnologia de sequenciação NGS. A análise que se pretende realizar tem por objetivos: caracterizar a variabilidade de cada um dos genomas individuais; relacionar a variabilidade existente entre os quatro genomas de indivíduos portugueses; relacionar a informação dos quatro genomas portugueses com a informação referente a populações caucasianas, bem como com outras populações representadas em projetos internacionais de sequenciação, como os 1000 Genomas e ESP; analisar a existência de variantes pontuais associadas a fenótipo; determinar a ancestralidade genética global dos quatro genomas de indivíduos portugueses; e caracterizar elementos genéticos reguladores que são transcritos e não são traduzidos, nomeadamente os miRNAs, inferindo o seu possível significado funcional.

Material e Métodos

1. Caracterização dos indivíduos em estudo

Foram obtidas amostras de DNA genómico a partir de quatro indivíduos adultos, do género masculino e, até à data da recolha das amostras, de estado autodeclarado saudável. Os quatro indivíduos apresentam nacionalidade e etnia autodeclarada Portuguesa. Foi dado consentimento informado para o estudo, pelos quatro indivíduos.

O estudo respeita os princípios da Declaração de Helsínquia, assegurando a confidencialidade dos dados obtidos assim como as boas práticas laboratoriais.

2. Sequenciação dos genomas humanos

2.1. Extração de DNA

Uma amostra de sangue total de cada indivíduo em estudo foi colhida para um tubo com ácido etilenodiaminotetracético (EDTA), a partir da qual se extraiu o DNA genómico através do Kit QIAamp DNA Blood Mini (Qiagen). A pureza do DNA foi avaliada por eletroforese em gel de agarose 0,8% e a quantidade medida pelo Nanodrop (NanoDrop Technologies). As amostras foram concentradas numa coluna Amicon (Millipore) de modo a ser obtida a concentração necessária para a sequenciação ($\geq$50ng/µl).

2.2. Preparação da biblioteca genómica e sequenciação

Cada biblioteca de sequenciação foi preparada a partir de 5 microgramas de DNA puro. De forma resumida, o DNA genómico foi fragmentado de forma aleatória por fragmentação acústica no equipamento Covaris (Covaris). O DNA fragmentado para cada genoma foi processado pelo kit TrueSeq DNA Sample Preparation (Illumina) seguindo as instruções do fabricante. A sequenciação foi efetuada na plataforma HiSeq 2000 de acordo com as instruções do fabricante, na GenePool (Edinburgo, Escócia). Cada genoma foi sequenciado com a química HiSeq V3 (Illumina) em três regiões da placa de sequenciação (*flowcells*) e no formato *paired-end* 2X100 bp.

3. Análise bioinformática dos genomas sequenciados

3.1. *Pipeline* analítica dos genomas

3.1.1. Montagem dos genomas sequenciados

A qualidade das *reads* obtidas em cada um dos genomas sequenciados (H1, H2, H3 e H4) foi avaliada no programa FastQC, versão 0.10.1 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). O FastQC permitiu avaliar a qualidade da sequenciação relativamente a: qualidade associada à identificação de bases em cada posição; qualidade média das *reads*; comprimento das *reads*; proporção de distribuição dos diferentes tipos de bases nas *reads*; conteúdo em %G+C; e à possível ocorrência de conteúdo N.

As *reads* sequenciadas, correspondentes a cada genoma, foram então mapeadas contra o genoma humano de referência hg19 do UCSC pelo SOAP3-dp, utilizando os parâmetros *default* (125). Os potenciais duplicados resultantes da PCR, foram removidos pelo comando *Mark Duplicates* da plataforma Picard, versão 1.114 (http://picard.sourceforge.net). De seguida, na sequência consenso de cada um dos genomas resultante do alinhamento das respetivas *reads*, efetuou-se a identificação de variantes pontuais, SNPs e INDELs, através do programa *The Genome Analysis Toolkit* (GATK), versão 3.1 (126). Antecedendo imediatamente a identificação das variantes, o GATK realizou um realinhamento local e uma recalibração da qualidade em cada um dos ficheiros. Na análise GATK foi usado um limite mínimo para o score de confiança de Q30, juntamente com os parâmetros *default* do programa. As variantes identificadas foram, por fim, registadas em quatro ficheiros de formato VCF, consoante o genoma a que pertenciam.

3.1.2. Anotação das variantes identificadas

Os SNPs e INDELs identificados em cada um dos genomas em estudo foram anotados e determinada a possível consequência funcional de cada uma das variantes através da ferramenta bioinformática ANNOVAR (65). Os quatro ficheiros de formato VCF foram convertidos no formato de entrada do ANNOVAR, GFF3, e anotados de acordo com a anotação do genoma humano de referência hg19 do UCSC *Genome Browser*.

Para tal foram utilizadas as coordenadas das variantes anotadas e identificadas as regiões genómicas em que cada variante se localizava (Anexo 1). De seguida, o conjunto total das variantes de cada genoma foi comparado com o dbSNP, versão 137. Desta comparação resultou a anotação com um rs identificativo das variantes onde as posições e as alterações corresponderam a alterações anteriormente reportadas ao dbSNP137. A totalidade das variantes foi também comparada com a informação resultante dos projetos internacionais dos 1000 Genomas (população mundial) e ESP (população americana e europeia), para a anotação das MAF correspondentes a cada variante identificada (47, 54). As variantes anotadas para as regiões exónicas de cada genoma foram, de seguida, extraídas para a análise quer da sua ocorrência em zonas conservadas do genoma, quer para a aferição do seu potencial efeito funcional na proteína a que podem estar associadas.

3.1.2.1. Anotação funcional das variantes exónicas

Para prever o efeito das variantes codificantes na função da proteína foi utilizado, primeiramente, o *Sorting Intolerant from Tolerant algoritm* (SIFT) (127). Esta ferramenta bioinformática anota as variantes utilizando como parâmetro a homologia de sequências, calculando a probabilidade que uma substituição de aminoácido terá no efeito que exerce sobre a proteína. O SIFT categorizou as alterações em: D (*Damaging*), todas aquelas que o algoritmo prevê que afetem a função da proteína; e T (*Tolerated*) as variantes que se prevê serem toleradas ao nível do efeito funcional da proteína. De seguida, foi realizada uma nova anotação das variantes exónicas para prever o possível impacto da substituição na estabilidade e função das proteínas, recorrendo ao *Polymorphism Phenotyping* versão 2 (PolyPhen 2) (128), o qual usou como parâmetros de análise considerações estruturais, comparações evolutivas e anotações presentes na *Swiss-Prot* e na *UniRef100*. O PolyPhen 2 categorizou as variantes anotadas em: P (*Probably damaging*), o que significa que a alteração é prevista como deletéria com elevada confiança; D (*Possibly damaging*), que significa que a alteração é prevista como deletéria com baixa confiança; e com B (*Benign*), as alterações que são previstas como benignas com elevada confiança. Consecutivamente, para a análise das mesmas variantes exónicas foi utilizada uma aplicação bioinformática que permite determinar o *Likelihood Ratio Test* (LRT) (129). Pelo LRT foram anotadas as variantes que interrompem as sequências codificantes da proteína, nomeadamente aminoácidos altamente conservados, sendo excluídos deste algoritmo os parâmetros da

relação filogenética e a distância evolutiva entre as sequências. O LRT categorizou as variantes anotadas em: D (*Deleterious*), as variantes que cumprem todos os requerimentos do algoritmo para uma maior probabilidade de causar dano; em N (*Neutral*), as variantes com probabilidade de ter um efeito neutro na proteína; e em U (*Unknown*), as variantes que não detêm informação suficiente para o cálculo do LRT. Por fim, as variantes exónicas foram ainda anotadas pela aplicação *Mutation Taster* (130) cuja análise teve por base os parâmetros de conservação evolutiva, as alterações nos locais de *splicing*, a perda de características proteicas e as alterações que podem afetar o mRNA. O teste resultante da análise de cada alteração, segundo os parâmetros supracitados, foi analisado num *Naive Bayes Classifier*, que prevê o potencial da variante para conduzir à doença. Assim, o *Mutation Taster* classificou as alterações anotadas como: A (*Disease causing automatic*), todas as alterações determinadas por informação externa passíveis de serem deletérias; D (*Disease causing*), todas as alterações previstas pelo algoritmo de serem deletérias; P (*Polymorphism automatic*), as alterações determinadas por informação externa como tendo baixa probabilidade de causar dano; e em N (*Polymorphism*), as alterações previstas pelo algoritmo como tendo baixa probabilidade de causar dano.

Por último identificaram-se as variantes exónicas que, segundo a sua posição, estão em zonas consideradas conservadas do genoma, através do algoritmo *Phylogenetic P-value* (Phylop), que atribuiu um *score* a cada variante baseado no alinhamento múltiplo de 46 genomas (131).

3.2. Caracterização das variantes anotadas

Em cada genoma, da totalidade das variantes anotadas foram selecionadas as variantes que correspondiam aos seguintes critérios: variantes anotadas na região exónica; variantes que apresentaram a classificação D pelo algoritmo SIFT; a classificação D ou P pelo PolyPhen 2; a classificação D pelo LRT; e a classificação A ou D pelo *Mutation Taster*; foram removidas todas as variantes anotadas em zonas repetitivas do genoma. Foram aplicados os mesmos critérios para o conjunto das variantes que foram identificadas em comum nos quatro genomas analisados. O conjunto de variantes que resultou do processo de filtração para cada um dos genomas, assim como o conjunto das variantes partilhadas pelos quatro genomas, foi caracterizado a nível génico e funcional.

3.2.1. Caracterização de genes

Após a filtração e seleção do conjunto de variantes a caracterizar, recorreu-se ao *Search Tool for the Retrieval of Interacting Genes/Proteins* (STRING) versão 9.1 para a análise das interações proteína-proteína que decorrem dos genes afetados pelas variantes anteriormente selecionadas (132). O STRING realizou redes de interação dos produtos dos genes baseando-se em associações diretas (físicas) e indiretas (funcionais) das proteínas consideradas. O programa integra a informação de dados experimentais inseridos em bases de dados curadas (MINT; HPRD; BIND; DIP; BioGriD; KEGG; Reactome; IntAct; EcoCyc; NCI-Nature Pathway Interation Database; GO), a informação obtida por previsão computacional (através dos parâmetros: genes em proximidade; genes que codificam uma proteína de fusão; perfil filogenético; genes co-expressos na mesma condição; translação de interações verificadas noutros organismos), a informação de interações descritas na bibliografia (SGD; OMIM; todos os resumos da PubMed) e integra a informação das estruturas de proteínas (PBD), de forma a relatar uma interação específica e com significado entre o produto de 2 genes, que juntamente contribuem para o mesmo processo funcional. As redes de interação foram construídas a partir de um *score* mínimo de 0,4 para a confiança na associação, ou seja, um *score* considerado pelo programa de confiança média. O conjunto das variantes partilhadas pelos quatro genomas e filtradas pelas ferramentas bioinformáticas SIFT, PolyPhen 2, LRT e *Mutation Taster*, foram também caracterizadas para a análise das interações proteína-proteína que resultam dos genes com as variantes selecionadas, associações que estão descritas na base de dados IntAct e foram representadas pelo Cytoscape versão 3.1.1 (133).

Para caracterizar cada gene associado às variantes filtradas quanto ao processo biológico em que se envolve, bem como à função molecular inerente, foi utilizada a ferramenta bioinformática *Protein Analysis Through Evolutionary Relationships* (PHANTER) (134). A inserção da lista de genes no PHANTER para caracterização levou à anotação dos produtos desses genes via *Gene Ontology* (GO).

3.2.2. Caracterização das variantes associadas a doença

Foram analisadas as variantes filtradas de acordo com os parâmetros descritos em 3.2, para determinar eventuais associações, já descritas, destas alterações a fenótipos de

doença. Recorreu-se para esta caracterização ao *Ensembl Genome Browser* (135), o qual forneceu a informação de cada variante quanto ao seu envolvimento em efeitos fenotípicos, bem como quanto à prevalência destas na população mundial. Para esta caracterização, o *Ensembl* recorreu a dados do dbSNP137, da base de dados OMIM e dos projetos internacionais dos 1000 Genomas e ESP. As variantes associadas a fenótipo foram confirmadas manualmente, através da verificação da presença da alteração nas várias *reads* mapeadas contra a sequência de referência, pelo *software Integrative Genomic Viewer* (IGV) v2.1.9 (136).

3.2.3. Farmacogenómica

Realizou-se uma análise de marcadores farmacogenómicos em cada um dos genomas, avaliando as variantes filtradas de acordo com os critérios descritos em 3.2., na base de dados PharmGKB (137).

3.3. Caracterização de elementos reguladores não codificantes

Foi realizada a determinação dos SNPs com funções reguladoras (rSNP) ou SNPs em *linkage disequilibrium* (LD) com rSNPs, que se encontram na região intrónica de cada genoma, através da utilização da aplicação bioinformática rSNPBase, uma base de dados curada para rSNPs (http://rsnp.psych.ac.cn/) (138). Na análise efetuada através da rSNPBase foram identificados rSNPs referentes a três tipos de regulação: regulação transcricional proximal, cuja informação foi obtida a partir do *Ensembl* (135); regulação transcricional distal, com informação fornecida pelos dados do projeto ENCODE (56); e regulação pós-transcricional mediada por proteínas de ligação ao RNA, cuja informação é igualmente fornecida pelos dados do ENCODE. Já a identificação efetuada dos SNPs que estão em LD com rSNPs teve como base a informação presente nas fases I+II+III do projeto HapMap e na fase I dos 1000 Genomas (47, 48).

Foram ainda selecionados os SNPs de cada genoma cuja anotação referia a sua localização em miRNAs e caracterizados os miRNAs alterados com a informação funcional presente na miRTarBase versão 4.5 e Microcosm versão 5, com recurso ao programa Cytoscape versão 3.1.1 e à aplicação CyTargetLinker (139). As variantes anotadas nos miRNAs foram confirmadas manualmente, através da verificação da presença

da alteração nas várias *reads* mapeadas contra a sequência de referência, pelo *software* IGV, versão 2.1.9 (136).

3.4. Estudo da ancestralidade

A ancestralidade genética global de cada indivíduo português, cujo genoma foi sequenciado (H1, H2, H3 e H4), foi estimada pela utilização do LASER (77). Esta ferramenta bioinformática construiu um sistema de coordenadas de referência, pela aplicação da PCA, através da comparação de cada amostra referida com indivíduos de referência, cuja informação da ancestralidade é conhecida e um conjunto de cerca de 650.000 SNPs referente a cada um está disponível. Nesta análise, foram utilizados como referência os SNPs que constam do *Human Genome Diversity Panel* (HGDP), consistindo na informação de 938 indivíduos de 53 populações mundiais (140). Para caracterizar os genomas H1, H2, H3 e H4 quanto à ancestralidade genética relativamente ao modelo de referência, foram utilizados 626.165 SNPs do genoma H1, 632.898 SNPs do genoma H2, 632.859 SNPs do genoma H3 e 630.157 SNPs do genoma H4. Depois de construído o painel de referência resultante da PCA, foram analisadas a primeira componente principal (PC1) e segunda componente principal (PC2), que permitiram separar as populações continentais. Foram efetuadas três réplicas para a construção da PCA, tendo sido realizada a análise da ancestralidade com a média dos valores resultantes.

Resultados

1. Avaliação da qualidade da sequenciação

Quatro bibliotecas *paired-end* foram sequenciadas na plataforma HiSeq 2000, da Illumina. O comprimento das *reads* obtidas na sequenciação foi de 100 pb (Anexo 2). No total, ≈100 Gb foram geradas para cada genoma. Após a análise da qualidade, as *reads* dos genomas H1, H2, H3 e H4, foram mapeadas contra o genoma de referência hg19, tendo ≈ 98% das *reads* sido mapeadas numa única posição do genoma de referência. Em média, verificou-se uma cobertura de 34x para os genomas estudados (Tabela 2).

Tabela 2: Dados de mapeamento dos genomas H1, H2, H3 e H4, contra o genoma de referência hg19, pela utilização do programa SOAP3-dp.

Dados de mapeamento	**H1**	**H2**	**H3**	**H4**
Reads mapeadas (%)	97,17	97,95	98,08	97,79
Reads não mapeadas (%)	2,83	2,08	1,92	2,21
Cobertura	31x	34x	34x	35x

A qualidade das *reads* resultantes da sequenciação de cada um dos genomas foi analisada com a aplicação FastQC. Nos quatro genomas verificou-se em média um *Phred Score* de qualidade por base ≥ Q30 na maioria das posições das *reads* (Figura 6).

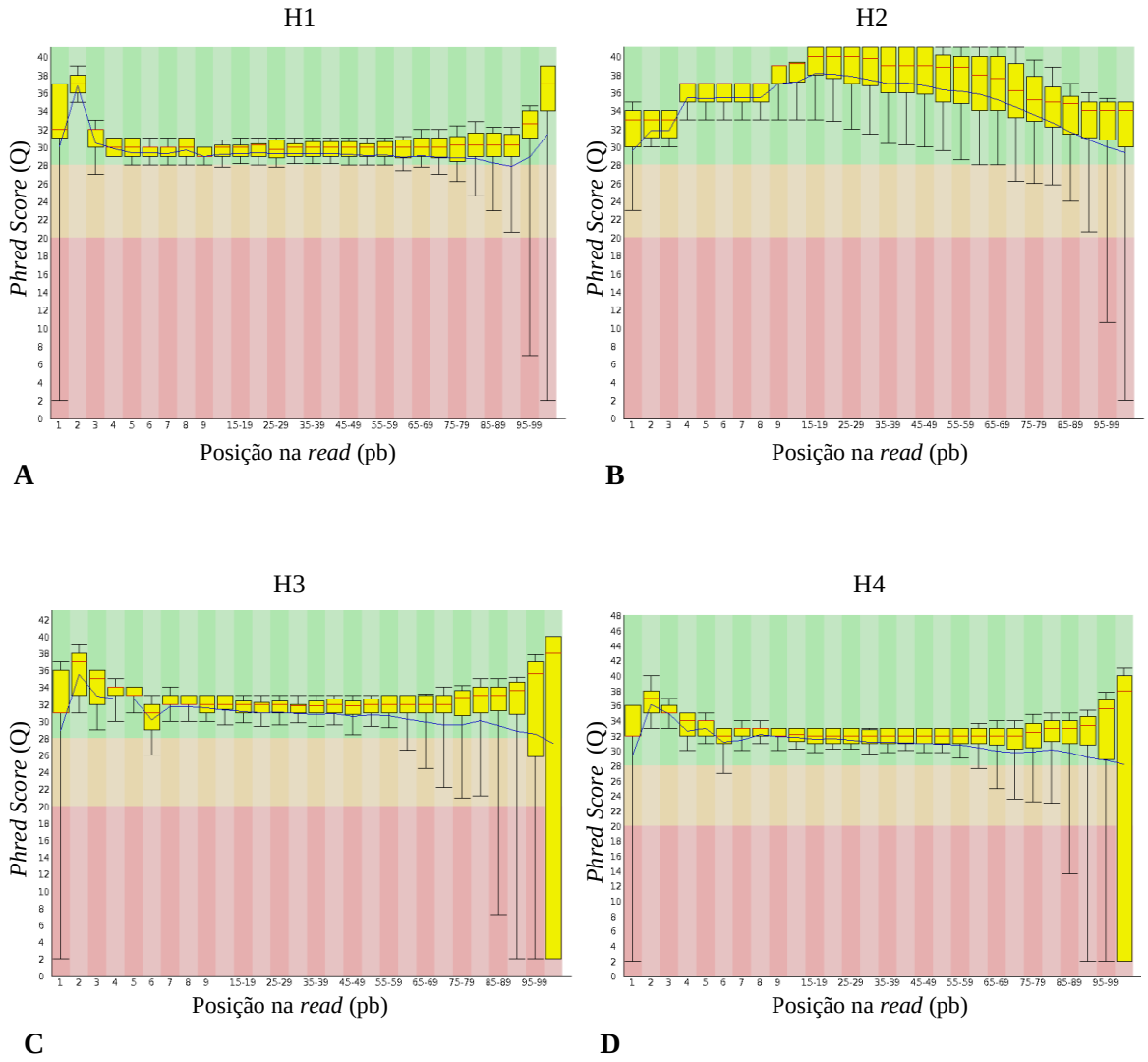


Figura 6: Análise dos valores de qualidade por base em cada posição das *reads* para o genoma H1 (A), o genoma H2 (B), o genoma H3 (C) e o genoma H4 (D), através da aplicação FastQC.

Do *score* de qualidade associado a cada posição de base na *read*, foi obtido o *score* médio da qualidade da r*ead* em cada genoma (Figura 7). O *Phred Score* calculado para o genoma H1 foi de Q30, para o genoma H2 foi de Q38, para os genomas H3 e H4 foi de Q32. Para analisar a existência de desvios derivados do enriquecimento de um ou mais tipos de bases, foi verificada a proporção média das bases em cada posição das *reads* (Anexo 3). Nos quatro genomas, não foi verificado qualquer enviesamento da proporção das bases nas *reads* sequenciadas, apresentando os conteúdos das quatro bases uma distribuição paralela.

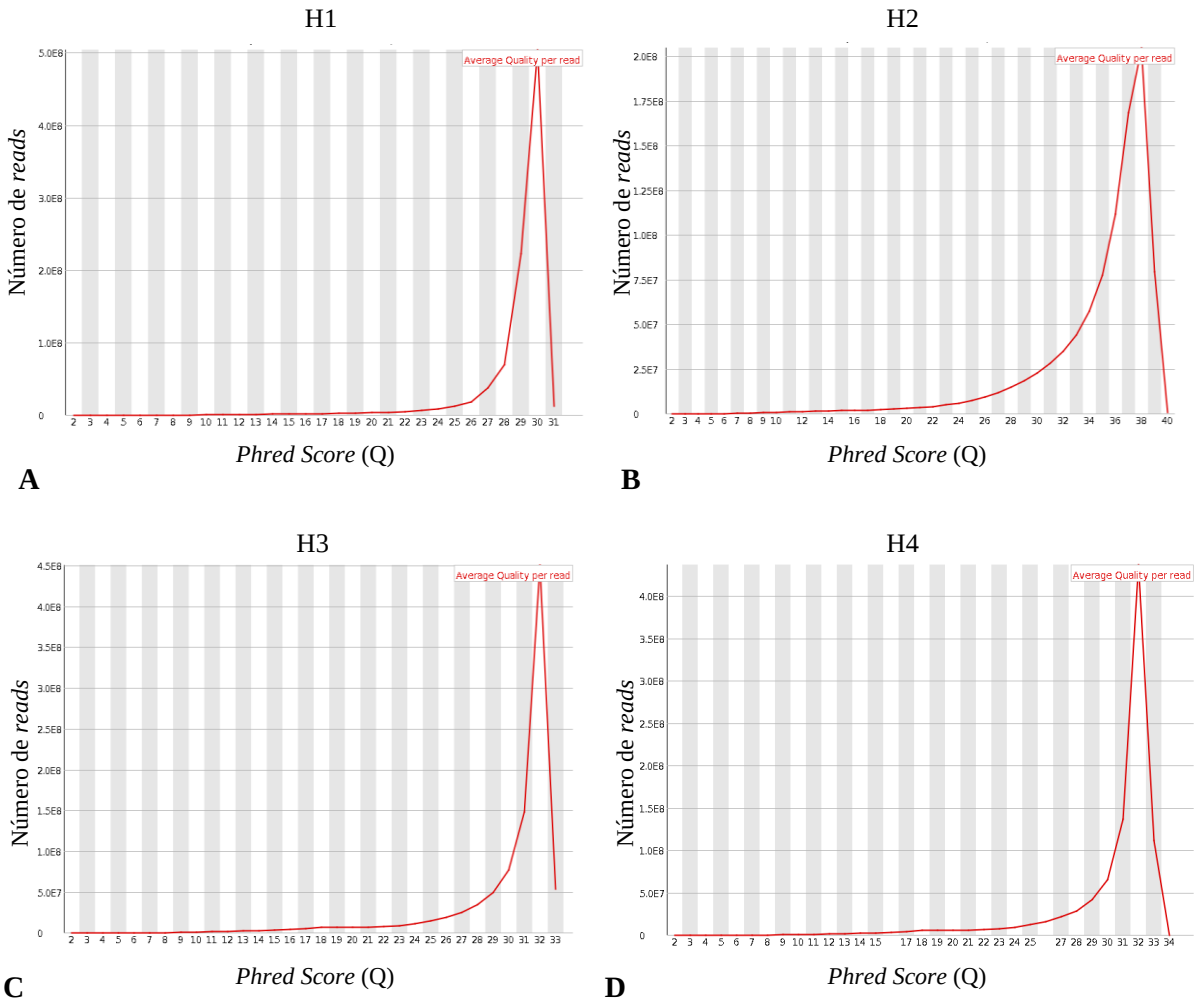


Figura 7: Análise da qualidade média (*Phred Score*) das *reads* obtidas na sequenciação do genoma H1 (A), genoma H2 (B), genoma H3 (C) e genoma H4 (D), através da aplicação FastQC.

O conteúdo em %G+C de cada genoma foi, respetivamente, 45% para o genoma H1, 43% para o H2, 44% para o H3 e 48% para o H4 (Anexo 4). Quando o sequenciador foi incapaz de identificar a base com confiança suficiente, atribuiu a essa posição um conteúdo N, no que deveria ser uma base convencional. Analisando o conteúdo N nos quatro genomas, não foi verificada a sua ocorrência em nenhum deles.

2. Análise dos genomas sequenciados

2.1. Distribuição das variantes por cromossoma

No início da análise dos quatro genomas sequenciados, H1, H2, H3 e H4, determinou-se o número de variantes pontuais, SNPs e INDELs, de cada um, assim como a sua distribuição pelos cromossomas nucleares.

No genoma H1 foram identificados 3.826.898 SNPs e 353.768 INDELs, perfazendo um total de 4.180.666 variantes pontuais detetadas. Os valores determinados distribuíram-se por todos os cromossomas, tendo apresentado o cromossoma 2 o maior número de variantes, 322.193, seguido pelo cromossoma 1 com 311.867 variantes. Já o cromossoma Y apresentou o menor número de variantes, com um total de 57.550 (Figura 8). O genoma H2 apresentou um total de 4.935.137 variantes pontuais, repartidas em 4.398.887 SNPs e 536.250 INDELs. A totalidade das variantes distribuiu-se pelos 46 cromossomas, tendo-se verificado que o cromossoma 2 e o cromossoma 1 apresentaram o maior número de alterações, 381.945 e 365.987, respetivamente. O cromossoma Y apresentou o menor número de alterações, 57.933 (Figura 9). No genoma H3, após a identificação das variantes pontuais, obteve-se um valor de 4.350.872 SNPs e 503.832 INDELs, perfazendo um total de 4.854.704 variantes pontuais. Os cromossomas 2 e 1 apresentaram os valores de alterações mais elevados, 370.673 e 357.419, respetivamente, e o cromossoma Y apresentou o valor mais baixo de alterações, 53.869 (Figura 10). No genoma H4 verificou-se a existência de 4.376.224 variantes pontuais, divididas entre 4.010.775 SNPs e 365.449 INDELs. A distribuição das variantes supramencionadas do genoma H4 ocorreu pelos cromossomas nucleares, tendo-se verificado que o cromossoma 2 apresentou o maior número de SNPs e INDELs, cerca de 334.348 no total, seguindo-se o cromossoma 1 com 323.064. A par dos restantes genomas, o cromossoma Y do genoma H4 também apresentou o menor número de alterações, 52.923 (Figura 11).

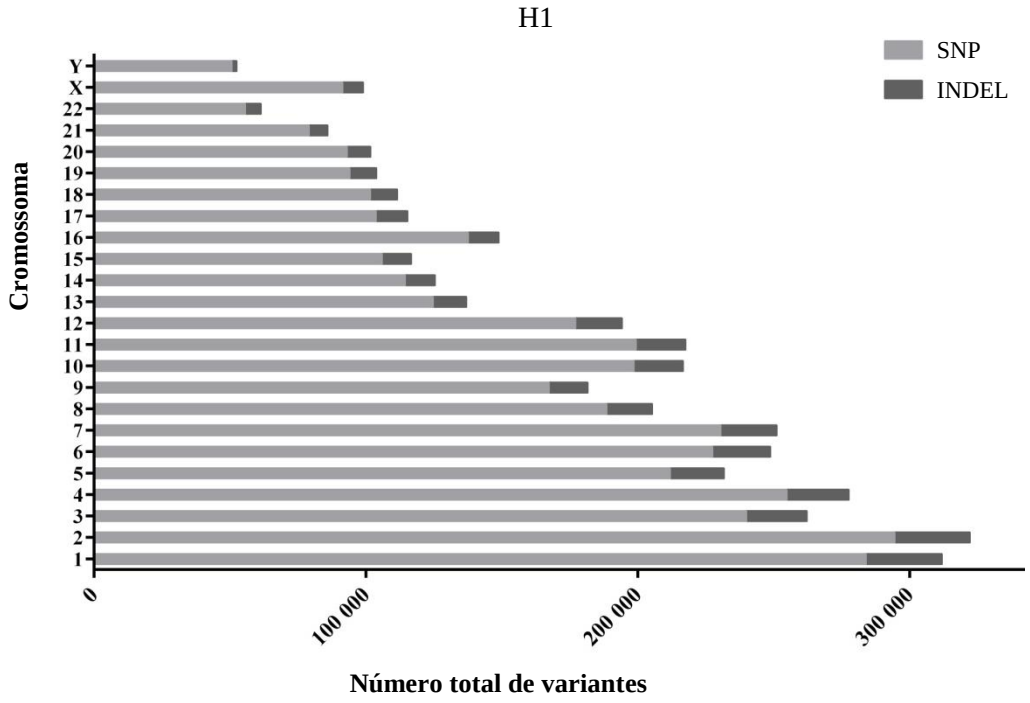


Figura 8: Distribuição dos SNPs e INDELs pelos cromossomas nucleares, identificados no genoma H1.

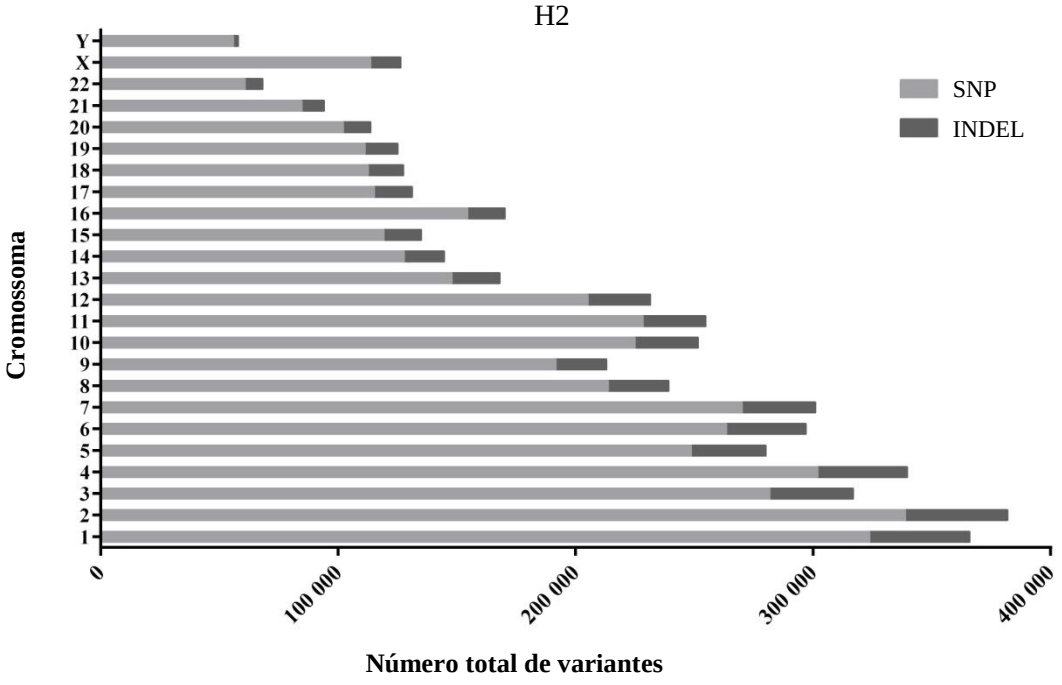


Figura 9: Distribuição dos SNPs e INDELs pelos cromossomas nucleares, identificados no genoma H2.

H3

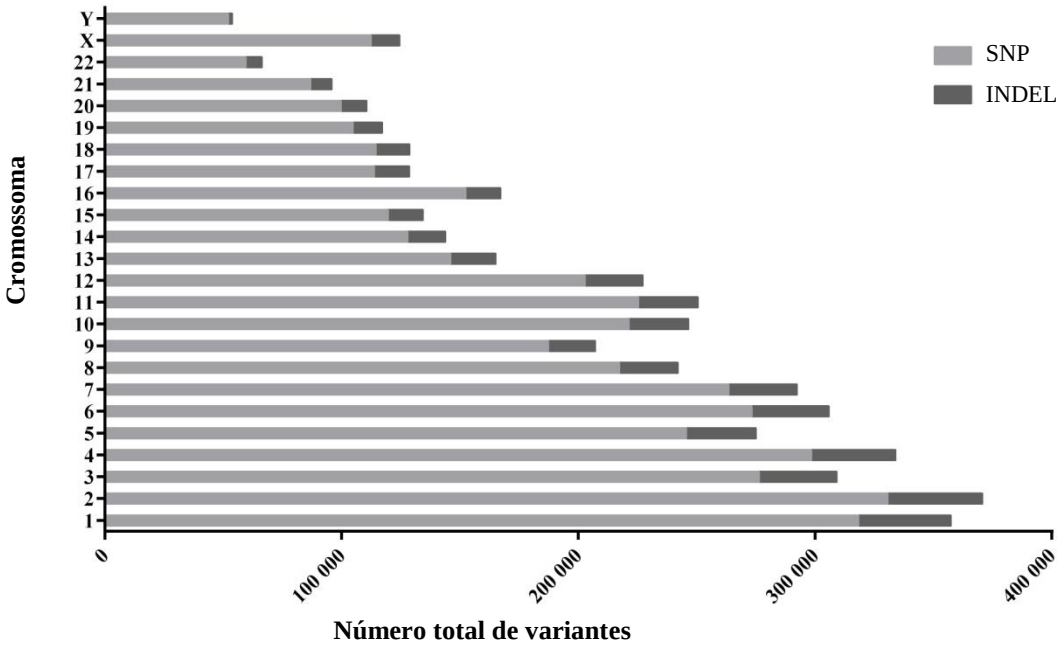


Figura 10: Distribuição dos SNPs e INDELs pelos cromossomas nucleares, identificados no genoma H3.

H4

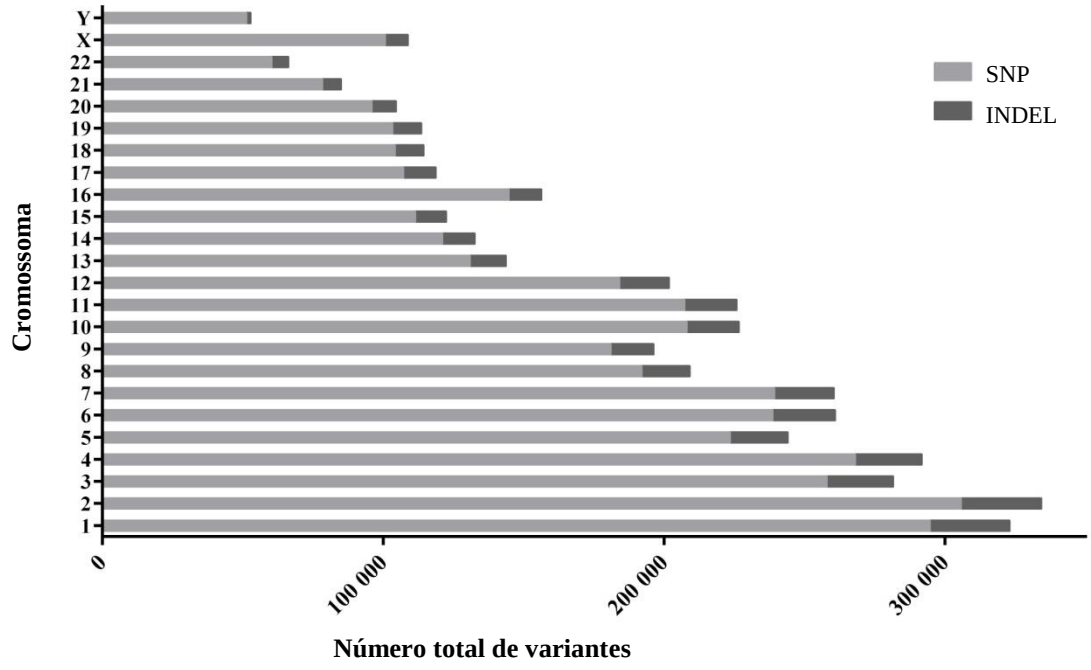


Figura 11: Distribuição dos SNPs e INDELs pelos cromossomas nucleares, identificados no genoma H4.

Atendendo ao número de pb que compõe cada cromossoma, foi calculado o rácio de bases alteradas relativamente ao número de pb por cromossoma, em cada genoma. No genoma H1, o cromossoma 19 e o cromossoma 21 apresentaram a maior taxa de variação, cada um com o valor de 0,16%, tendo os cromossomas X, Y e 15 apresentado os valores percentuais mais baixos, com 0,06%, 0,09% e 0,10%, respetivamente. O genoma H1 apresentou uma taxa de variação média de 0,12% (Tabela 3). Nos valores relativos ao rácio de variação de cada cromossoma no genoma H2, os cromossomas 19 e 21 apresentaram os valores mais elevados, 0,19% e 0,18%, respetivamente, enquanto o cromossoma X apresentou o valor mais baixo, 0,07%. O genoma H2 apresentou um valor médio de variabilidade de 0,14% (Tabela 3). No genoma H3, relativamente à taxa de variação por cromossoma foram os cromossomas 19 e 21 os que apresentaram os valores de variação mais elevados, ambos com 0,18%. Neste genoma, o cromossoma X voltou a apresentar a menor taxa de variação com 0,07%. A taxa de variação deste genoma teve o valor de 0,14% (Tabela 4). Considerando o genoma H4, foi o cromossoma 19 que apresentou a maior taxa de variação, com 0,17%, sendo o cromossoma X aquele cuja taxa de variação foi menor, com 0,06%. O genoma H4 apresentou uma taxa de variabilidade média de 0,13% (Tabela 4).

.

Tabela 3: Distribuição de SNPs e INDELs nos genomas H1 e H2. Análise do rácio de bases variantes, relativamente ao número de SNPs em cada cromossoma.

		H1				H2			
	Comprimento (pb)	**SNPs**	**Rácio de bases variantes (%)**	**INDELs**	**Total de alterações**	**SNPs**	**Rácio de bases variantes (%)**	**INDELs**	**Total de alterações**
Cromossoma 1	**249.250.621**	283.995	0,11	27.872	311.867	323.849	0,13	42.138	365.987
Cromossoma 2	**243.199.373**	294.547	0,12	27.646	322.193	338.839	0,14	43.106	381.945
Cromossoma 3	**198.022.430**	239.926	0,12	22.347	262.273	281.741	0,14	35.109	316.850
Cromossoma 4	**191.154.276**	254.859	0,13	22.748	277.607	301.920	0,16	37.775	339.695
Cromossoma 5	**180.915.260**	211.935	0,12	19.834	231.769	248.709	0,14	31.377	280.086
Cromossoma 6	**171.115.067**	227.515	0,13	21.261	248.776	263.548	0,15	33.441	296.989
Cromossoma 7	**159.138.663**	230.528	0,14	20.588	251.116	270.075	0,17	30.962	301.037
Cromossoma 8	**146.364.022**	188.589	0,13	16.749	205.338	213.740	0,15	25.464	239.204
Cromossoma 9	**141.213.431**	167.421	0,12	14.223	181.644	191.721	0,14	21.302	213.023
Cromossoma 10	**135.534.747**	198.605	0,15	18.083	216.688	225.004	0,17	26.521	251.525
Cromossoma 11	**135.006.516**	199.402	0,15	18.178	217.580	228.365	0,17	26.407	254.772
Cromossoma 12	**133.851.895**	177.160	0,13	17.074	194.234	205.235	0,15	26.274	231.509
Cromossoma 13	**115.169.878**	124.731	0,11	12.254	136.985	147.888	0,13	20.224	168.112
Cromossoma 14	**107.349.540**	114.435	0,11	11.059	125.494	127.794	0,12	16.913	144.707
Cromossoma 15	**102.531.392**	105.998	0,10	10.681	116.679	119.314	0,12	15.762	135.076
Cromossoma 16	**90.354.753**	137.606	0,15	11.262	148.868	154.509	0,17	15.741	170.250
Cromossoma 17	**81.195.210**	103.725	0,13	11.638	115.363	115.282	0,14	15.957	131.239
Cromossoma 18	**78.077.248**	101.678	0,13	9.879	111.557	112.661	0,14	14.861	127.522
Cromossoma 19	**59.128.983**	94.066	0,16	9.822	103.888	111.354	0,19	13.837	125.191
Cromossoma 20	**63.025.520**	93.120	0,15	8.648	101.768	102.119	0,17	11.591	113.710
Cromossoma 21	**48.129.895**	79.156	0,16	6.799	85.955	84.787	0,18	9.315	94.102
Cromossoma 22	**51.304.566**	55.672	0,11	5.752	61.424	60.868	0,12	7.418	68.286
Cromossoma X	**155.270.560**	91.498	0,06	7.552	99.050	113.703	0,07	12.684	126.387
Cromossoma Y	**59.373.566**	50.731	0,09	1.819	57.550	55.862	0,09	2.071	57.933
Total:	**3.095.677.412**	**3.826.898**	**0,12**	**353.768**	**4.180.666**	**4.398.887**	**0,14**	**536.250**	**4.935.137**

Tabela 4: Distribuição de SNPs e INDELs nos genomas H3 e H4. Análise do rácio de bases variantes, relativamente ao número de SNPs em cada cromossoma.

		H3				H4			
	Comprimento (pb)	**SNPs**	**Rácio de bases variantes (%)**	**INDELs**	**Total de alterações**	**SNPs**	**Rácio de bases variantes (%)**	**INDELs**	**Total de alterações**
Cromossoma 1	**249.250.621**	318.319	0,13	39.100	357.419	294.727	0,12	28.337	323.064
Cromossoma 2	**243.199.373**	330.748	0,14	39.925	370.673	305.821	0,13	28.527	334.348
Cromossoma 3	**198.022.430**	276.372	0,14	32.814	309.186	257.966	0,13	23.614	281.580
Cromossoma 4	**191.154.276**	298.425	0,16	35.586	334.011	268.056	0,14	23.774	291.830
Cromossoma 5	**180.915.260**	245.644	0,14	29.389	275.033	223.542	0,12	20.567	244.109
Cromossoma 6	**171.115.067**	273.280	0,16	32.581	305.861	238.667	0,14	22.320	260.987
Cromossoma 7	**159.138.663**	263.504	0,17	28.851	292.355	239.319	0,15	21.154	260.473
Cromossoma 8	**146.364.022**	217.371	0,15	24.702	242.073	192.073	0,13	17.154	209.227
Cromossoma 9	**141.213.431**	187.372	0,13	19.787	207.159	181.125	0,13	15.231	196.356
Cromossoma 10	**135.534.747**	221.425	0,16	25.080	246.505	208.175	0,15	18.545	226.720
Cromossoma 11	**135.006.516**	225.397	0,17	25.110	250.507	207.283	0,15	18.571	225.854
Cromossoma 12	**133.851.895**	202.869	0,15	24.393	227.262	184.194	0,14	17.605	201.799
Cromossoma 13	**115.169.878**	146.069	0,13	19.060	165.129	130.919	0,11	12.812	143.731
Cromossoma 14	**107.349.540**	127.927	0,12	15.977	143.904	121.093	0,11	11.545	132.638
Cromossoma 15	**102.531.392**	119.702	0,12	14.760	134.462	111.450	0,11	11.024	122.474
Cromossoma 16	**90.354.753**	152.398	0,17	14.779	167.177	144.778	0,16	11.564	156.342
Cromossoma 17	**81.195.210**	113.788	0,14	14.897	128.685	107.224	0,13	11.537	118.761
Cromossoma 18	**78.077.248**	114.515	0,15	14.267	128.782	104.235	0,13	10.199	114.434
Cromossoma 19	**59.128.983**	104.844	0,18	12.362	117.206	103.388	0,17	10.254	113.642
Cromossoma 20	**63.025.520**	99.776	0,16	10.855	110.631	95.987	0,15	8.687	104.674
Cromossoma 21	**48.129.895**	86.894	0,18	8.955	95.849	78.370	0,16	6.733	85.103
Cromossoma 22	**51.304.566**	59.626	0,12	6.810	66.436	60.319	0,12	5.985	66.304
Cromossoma X	**155.270.560**	112.448	0,07	12.082	124.530	100.723	0,06	8.128	108.851
Cromossoma Y	**59.373.566**	52.159	0,09	1.710	53.869	51.341	0,09	1.582	52.923
Total:	**3.095.677.412**	**4.350.872**	**0,14**	**503.832**	**4.854.704**	**4.010.775**	**0,13**	**365.449**	**4.376.224**

2.2. Distribuição das variantes por região genómica

Após terem sido mapeados os 4 genomas contra o genoma de referência hg19, os SNPs e INDELs identificados foram anotados nas diferentes regiões genómicas (Anexo 1).

O genoma H1 apresentou uma distribuição de SNPs em que 2.466.069 (64,4%) são referentes à região intergénica e 1.360.829 SNPs (35,6%) à região intragénica. Dos 1.360.829 SNPs anotados na região intragénica, 24.189 (0,6%) foram identificados na região exónica e 1.173.403 (30,7%) na região intrónica. Dos restantes SNPs encontrados na região intragénica destacaram-se os presentes na região 5'UTR e na região 3'UTR, com 4.060 (0,1%) e 22.116 (0,6%), respetivamente; os SNPs presentes na região *upstream* e na região *downstream*, com 21.349 (0,6%) e 21.998 (0,6%), respetivamente; e os SNPs presentes na região ncRNA intrónica e na região ncRNA exónica, com 84.421 (2,2%) e 8.246 (0,2%), respetivamente (Figura 12).

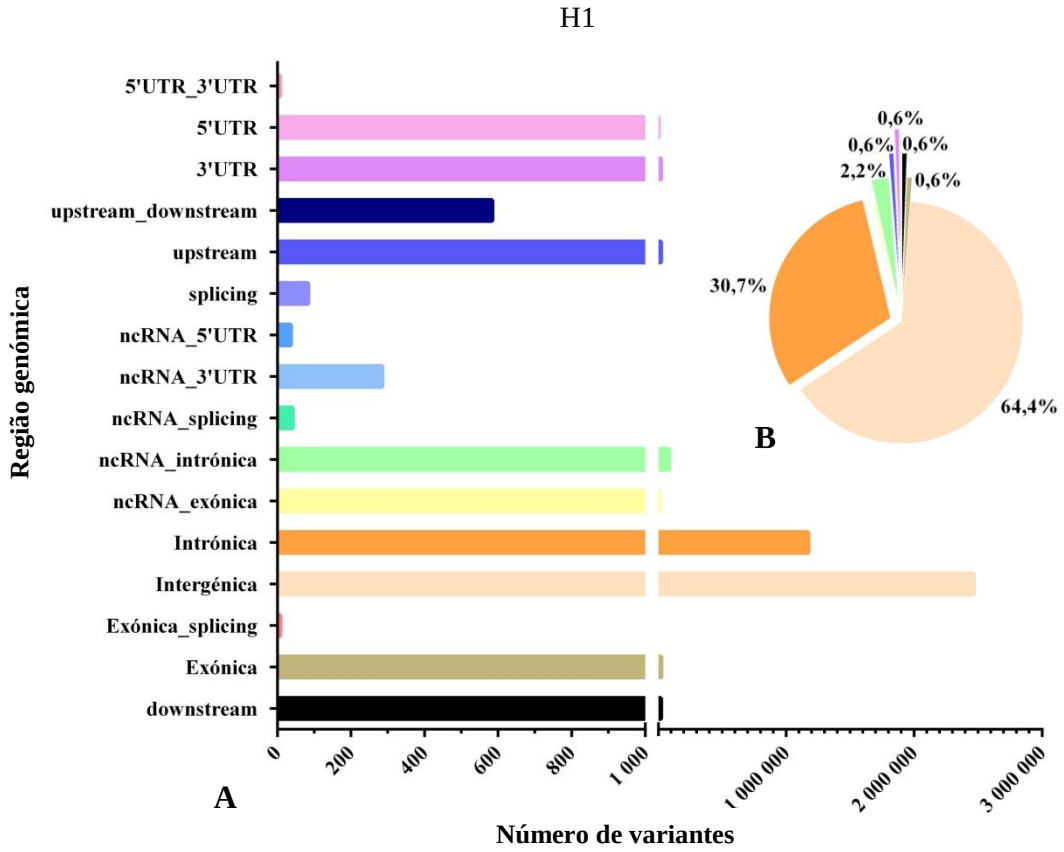


Figura 12: Distribuição dos SNPs pelas regiões genómicas do genoma H1 (A) e valores percentuais respetivos das regiões com maior abundância de SNPs (B).

Relativamente aos INDELs identificados no genoma H1, 209.599 (59,2%) encontravam-se na região intergénica e 144.169 (40,8%) na região intragénica. Dos 144.169 INDELs identificados na região intragénica, 466 (0,1%) estavam na região exónica e 126.541 (35,8%) na região intrónica. Destacaram-se ainda os INDELs presentes na região 3'UTR com 2.976 (0,8%); na região *upstream* e na região *downstream*, com 2.477 (0,7%) e 2.769 (0,8%), respetivamente; e na região ncRNA intrónica com 7.830 (2,2%) (Figura 13).

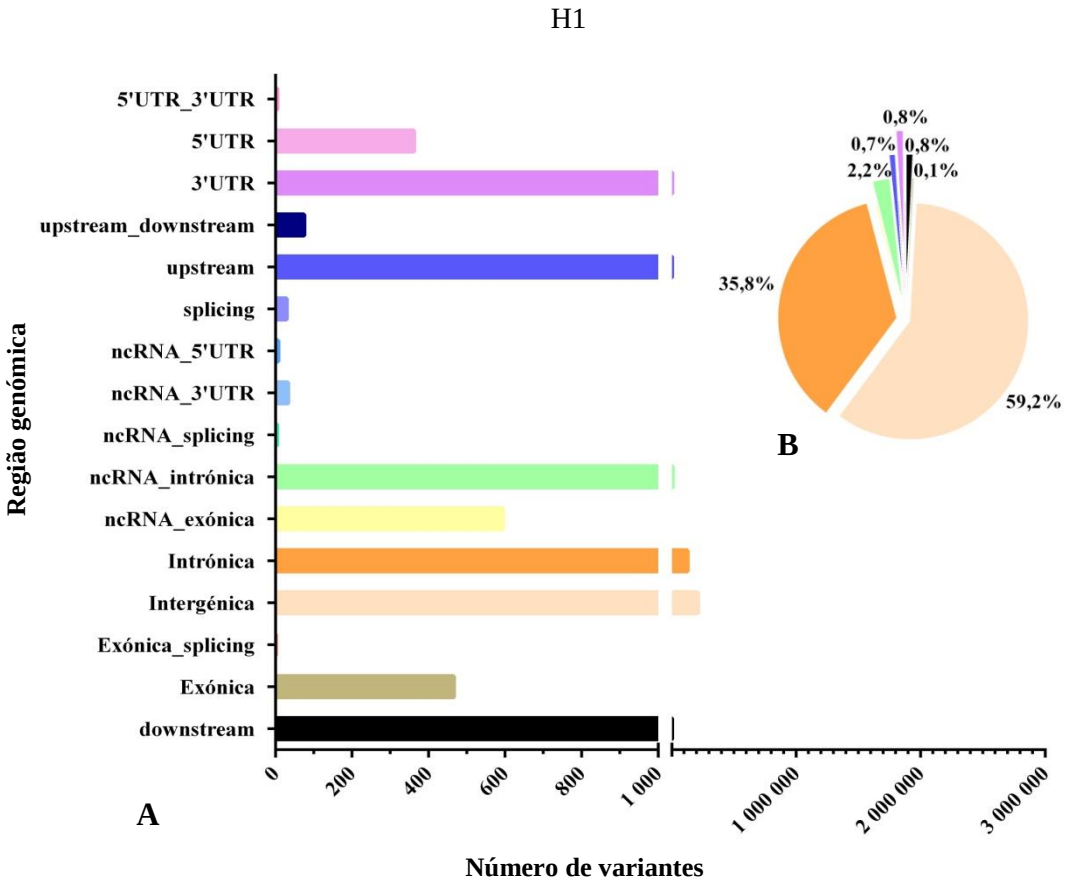


Figura 13: Distribuição dos INDELs pelas regiões genómicas do genoma H1 (A) e valores percentuais respetivos das regiões com maior abundância de INDELs (B).

O genoma H2 apresentou uma distribuição de SNPs com 2.851.545 (64,8%) presentes na região intergénica e 1.547.342 (35,2%) presentes na região intragénica. Dos 1.547.342 SNPs anotados na região intragénica, 25.194 (0,6%) foram identificados na região exónica e 1.340.139 (30,5%) na região intrónica. Dos restantes SNPs anotados na

região intragénica destacaram-se os presentes na região 5'UTR e na região 3'UTR, com 4.925 (0,1%) e 25.076 (0,6%), respetivamente; os SNPs presentes na região *upstream* e na região *downstream*, com 23.954 (0,5%) e 24.044 (0,5%), respetivamente; e os SNPs presentes na região ncRNA intrónica e na região ncRNA exónica, com 94.064 (2,1%) e 8.762 (0,2%), respetivamente (Figura 14).

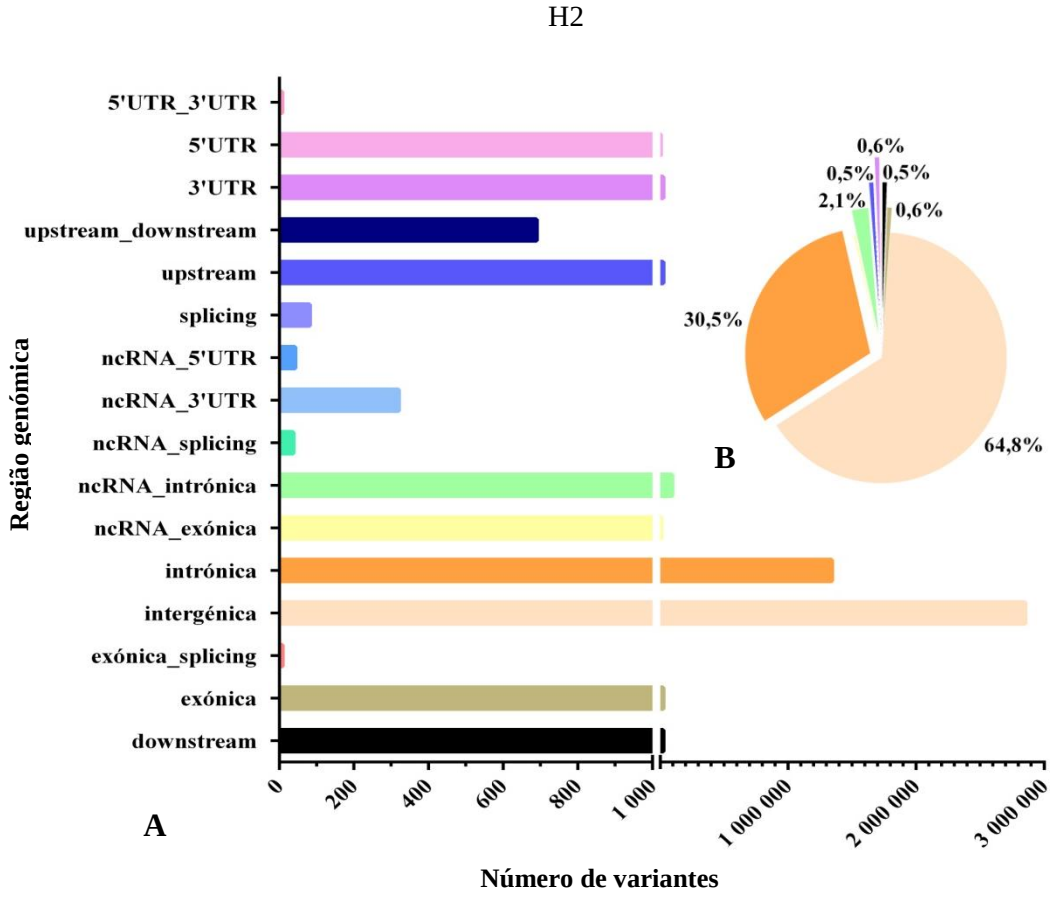


Figura 14: Distribuição dos SNPs pelas regiões genómicas do genoma H2 (A) e valores percentuais respetivos das regiões com maior abundância de SNPs (B).

Em relação aos INDELs identificados no genoma H2, 319.959 (59,7%) estavam presentes na região intergénica e 216.291 (40,3%) na região intragénica. Dos 216.291 INDELs presentes na região intragénica, 528 (0,1%) encontravam-se na região exónica e 191.234 (35,7%) na região intrónica. Destacaram-se ainda os INDELs presentes na região 3'UTR com 4.435 (0,8%); na região *upstream* e na região *downstream*, com 3.401 (0,6%)

e 3.841 (0,7%), respetivamente; e na região ncRNA intrónica com 11.363 (2,1%) (Figura 15).

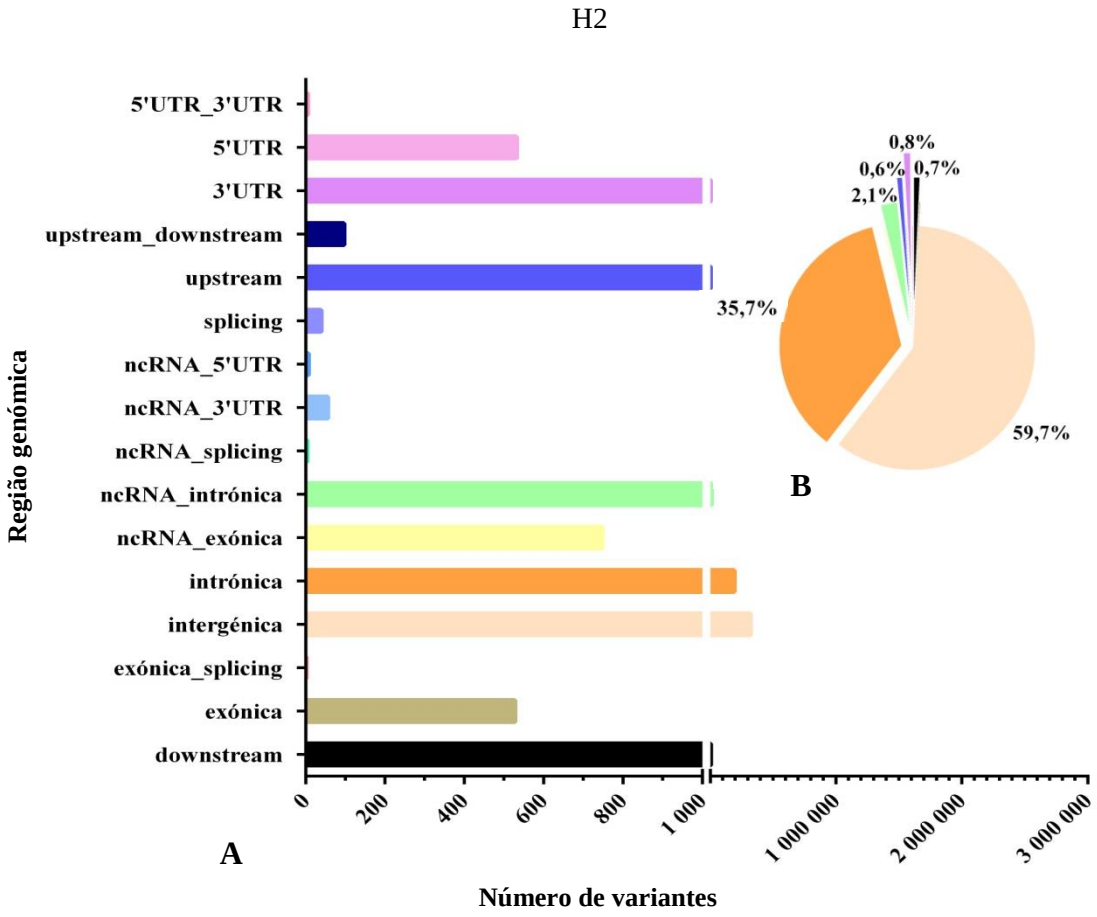


Figura 15: Distribuição dos INDELs pelas regiões genómicas do genoma H2 (A) e valores percentuais respetivos das regiões com maior abundância de INDELs (B).

O genoma H3 apresentou uma distribuição de SNPs em que 2.815.966 (64,7%) se localizavam na região intergénica e 1.534.906 (35,3%) na região intragénica. Dos 1.534.906 SNPs anotados na região intragénica, 25.267 (0,6%) são referentes à região exónica e 1.327.090 (30,5%) à região intrónica. Dos restantes SNPs presentes na região intragénica destacaram-se os presentes na região 5'UTR e na região 3'UTR, com 4.796 (0,1%) e 24.677 (0,6%), respetivamente; os SNPs presentes na região *upstream* e na região *downstream*, com 23.667 (0,5%) e 24.003 (0,6%), respetivamente; e os SNPs presentes na região ncRNA intrónica e na região ncRNA exónica, com 95.348 (2,2%) e 8.920 (0,2%), respetivamente (Figura 16).

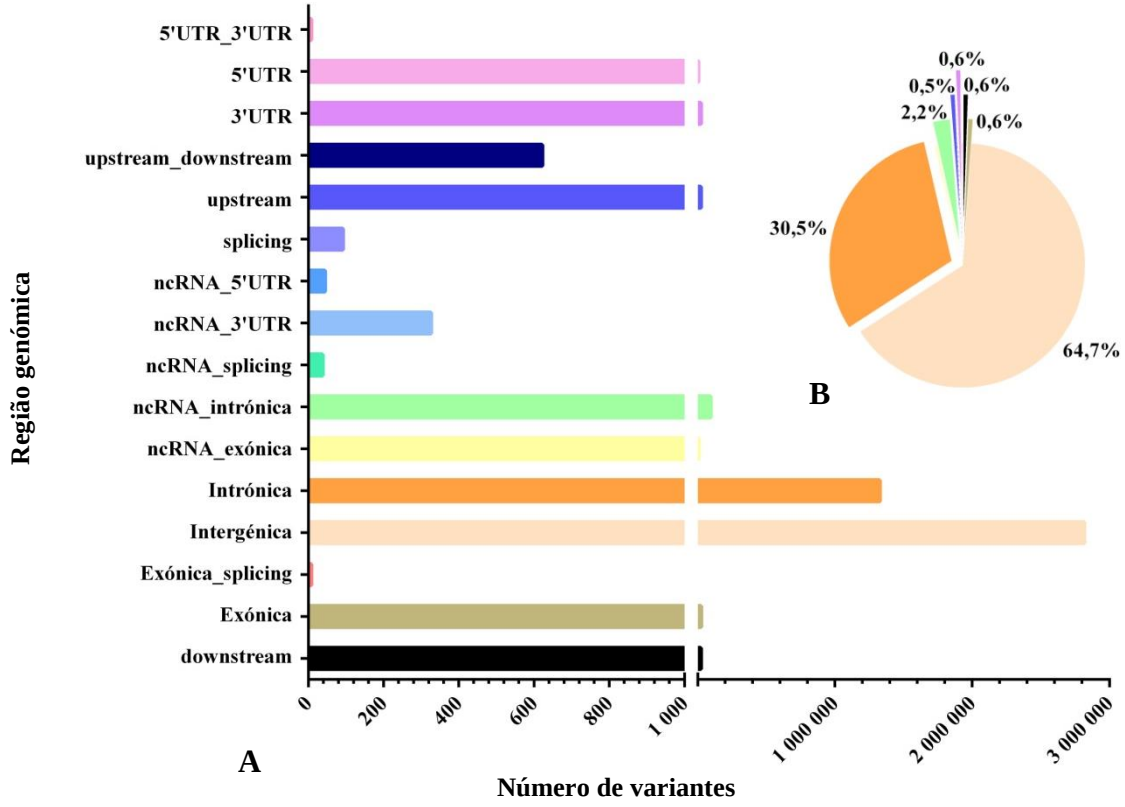


Figura 16: Distribuição dos SNPs pelas regiões genómicas do genoma H3 (A) e valores percentuais respetivos das regiões com maior abundância de SNPs (B).

Relativamente aos INDELs identificados no genoma H3, 301.291 (59,8%) estavam presentes na região intergénica e 202.541 (40,2%) na região intragénica. Dos 202.541 INDELs identificados na região intragénica, 534 (0,1%) observaram-se na região exónica e 178.586 (35,4%) na região intrónica. Destacaram-se ainda os INDELs presentes na região 3'UTR com 4.134 (0,8%); na região *upstream* e na região *downstream*, com 3.228 (0,6%) e 3.593 (0,7%), respetivamente; e na região ncRNA intrónica com 11.059 (2,2%) (Figura 17).

H3

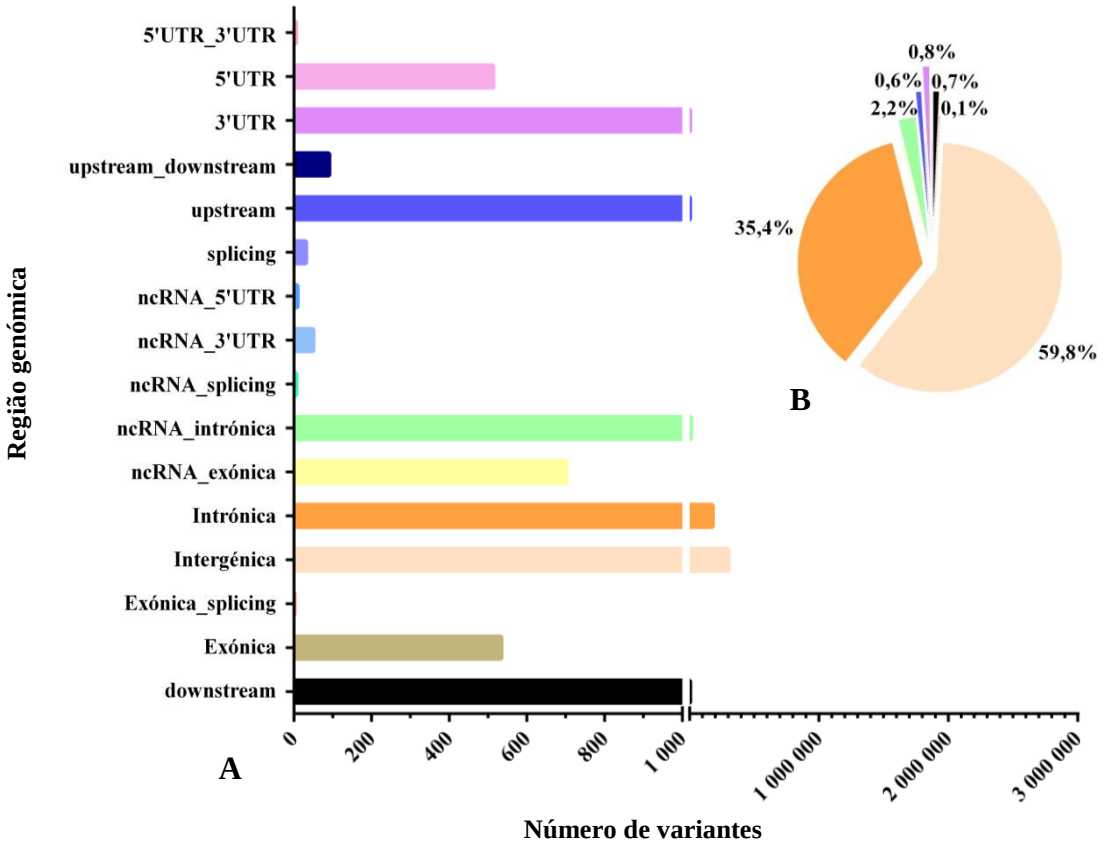


Figura 17: Distribuição dos INDELs pelas regiões genómicas do genoma H3 (A) e valores percentuais respetivos das regiões com maior abundância de INDELs (B).

O genoma H4 apresentou uma distribuição dos seus SNPs em que 2.588.505 (64,5%) encontravam-se na região intergénica e 1.422.270 (35,5%) na região intragénica. Dos 1.422.270 SNPs anotados na região intragénica, 24.514 (0,6%) localizavam-se na região exónica e 1.227.308 (30,6%) na região intrónica. Dos restantes SNPs presentes na região intragénica destacaram-se os presentes na região 5'UTR e na região 3'UTR, com 4.707 (0,1%) e 22.339 (0,6%), respetivamente; os SNPs presentes na região *upstream* e na região *downstream,* com 22.643 (0,6%) e 22.614 (0,6%), respetivamente; e os SNPs presentes na região ncRNA intrónica e na região ncRNA exónica, com 88.449 (2,2%) e 8.558 (0,2%), respetivamente (Figura 18).

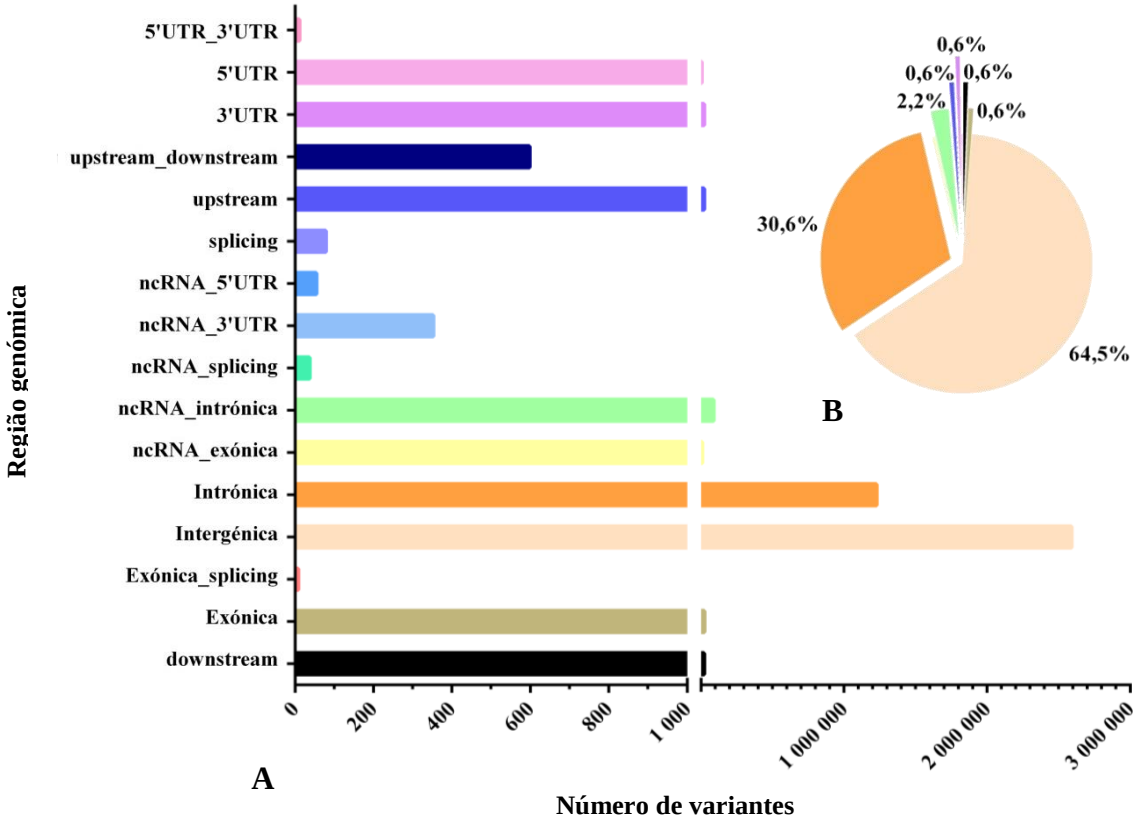


Figura 18: Distribuição dos SNPs pelas regiões genómicas do genoma H4 (A) e valores percentuais respetivos das regiões com maior abundância de SNPs (B).

Em relação aos INDELs identificados no genoma H4, 216.689 (59,3%) encontravam-se na região intergénica e 148.760 (40,7%) na região intragénica. Dos 148.760 INDELs presentes na região intragénica, 474 (0,1%) foram identificados na região exónica e 130.408 (35,7%) na região intrónica. Destacaram-se ainda os INDELs presentes na região 3'UTR com 2.919 (0,8%); na região *upstream* e na região *downstream*, com 2.754 (0,8%) e 2.822 (0,8%), respetivamente; e na região ncRNA intrónica com 8.099 (2,2%) (Figura 19).

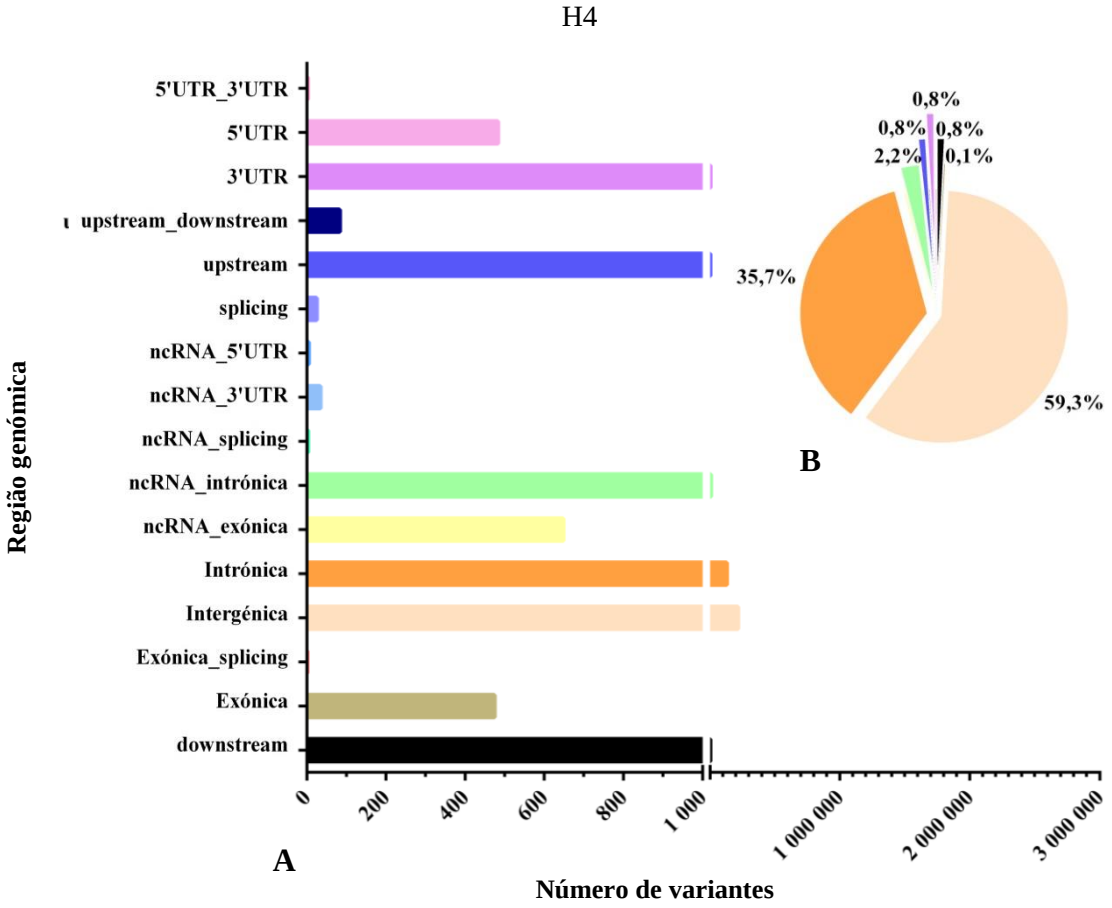


Figura 19: Distribuição dos INDELs pelas regiões genómicas do genoma H4 (A) e valores percentuais respetivos das regiões com maior abundância de INDELs (B).

2.3. Caracterização de SNPs

Os SNPs presentes nos quatro genomas sequenciados foram caracterizados relativamente à ocorrência de transições (Ts), ou seja a ocorrência da substituição de uma pirimidina, Citosina (C) ou Timina (T), por uma outra pirimidina, ou a substituição de uma purina, Adenina (A) ou Guanina (G), por uma outra purina; assim como relativamente à ocorrência de transversões (Tv), isto é a ocorrência da substituição de uma pirimidina por uma purina, ou o contrário.

No genoma H1, dos 3.826.898 SNPs identificados, 2.515.399 (65,7%) foram Ts e 1.311.499 (34,3%) Tv; no genoma H2, dos 4.398.887 SNPs identificados, 2.864.805 (65,1%) foram Ts e 1.534.082 (34,9%) Tv; no genoma H3, dos 4.350.872 SNPs identificados, 2.840.890 (65,3%) foram Ts e 1.509.982 (34,7%) foram Tv; no genoma H4,

dos 4.010.775 SNPs identificados, 2.630.905 (65,6%) foram Ts e 1.379.871 (34,4%) Tv. Nos quatro genomas estudados verificou-se, em todos os cromossomas, um número superior de Ts comparativamente ao número de Tv, sendo que nos primeiros cromossomas essa diferença foi mais acentuada comparativamente aos restantes cromossomas (Figura 20). Atendendo aos tipos de Ts, (T/C, C/T, G/A e A/G), os genomas H1, H2, H3 e H4 apresentaram uma ocorrência dos diferentes tipos na mesma proporção, ao nível de cada cromossoma (Figura 21). Efetuando a mesma análise para os diferentes tipos de Tv, (G/C, A/T, C/G, T/A, T/G, C/A, G/T e A/C), verificou-se que a ocorrência destes sucede numa proporção semelhante, ao nível de cada cromossoma nos quatro genomas estudados (Figura 22). Efetuou-se o cálculo do rácio Ts/Tv, tendo-se obtido o valor de 1,92 (2.515.399/1.311.499) para o genoma H1, 1,87 (2.864.805/1.534.082) para o genoma H2, 1,88 (2.840.890/1.509.982) para o genoma H3, e 1,91 (2.630.905/1.379.871) para o genoma H4.

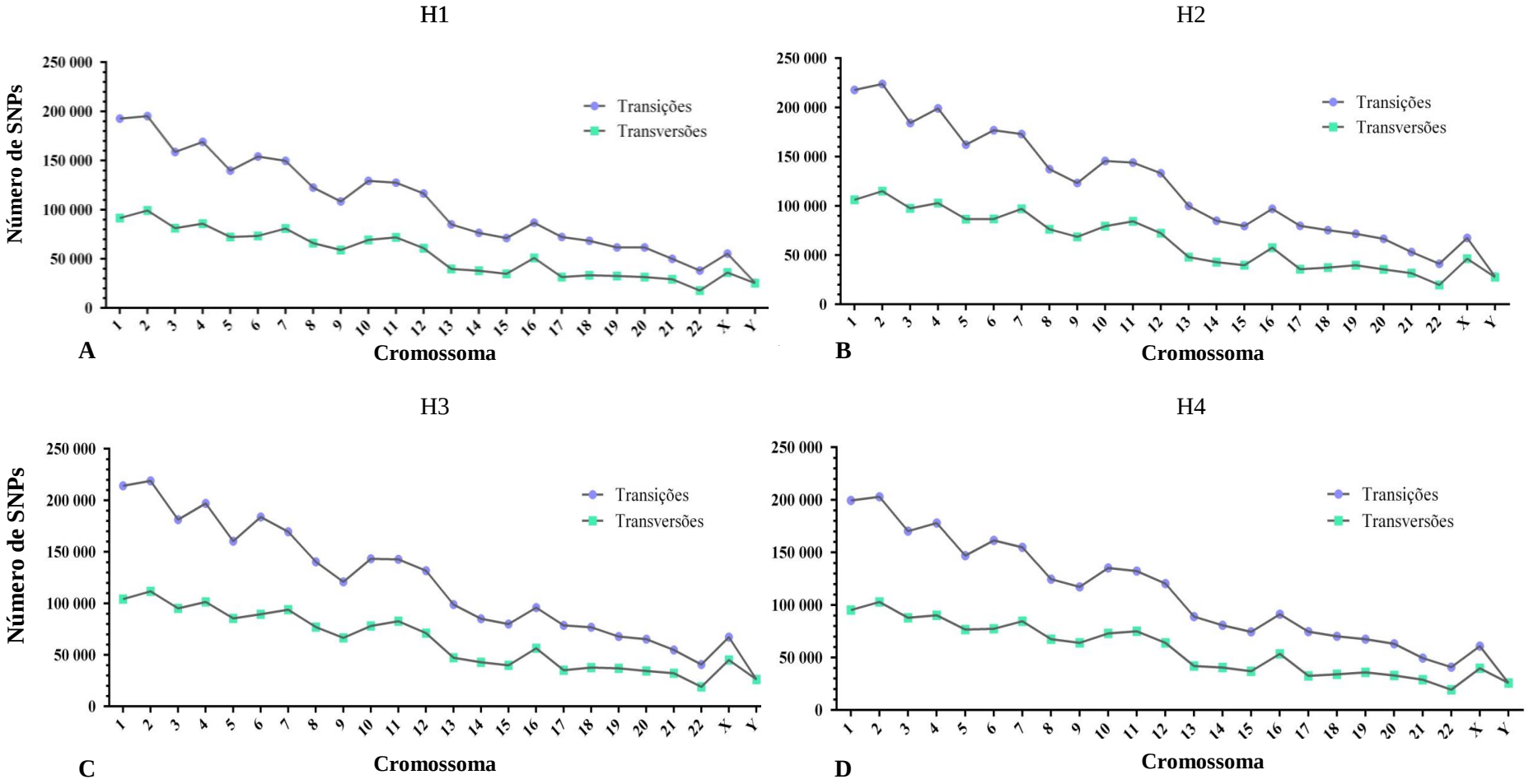


Figura 20: Distribuição de Ts e Tv pelos cromossomas nucleares do genoma H1 (A), genoma H2 (B), genoma H3 (C) e genoma H4 (D).

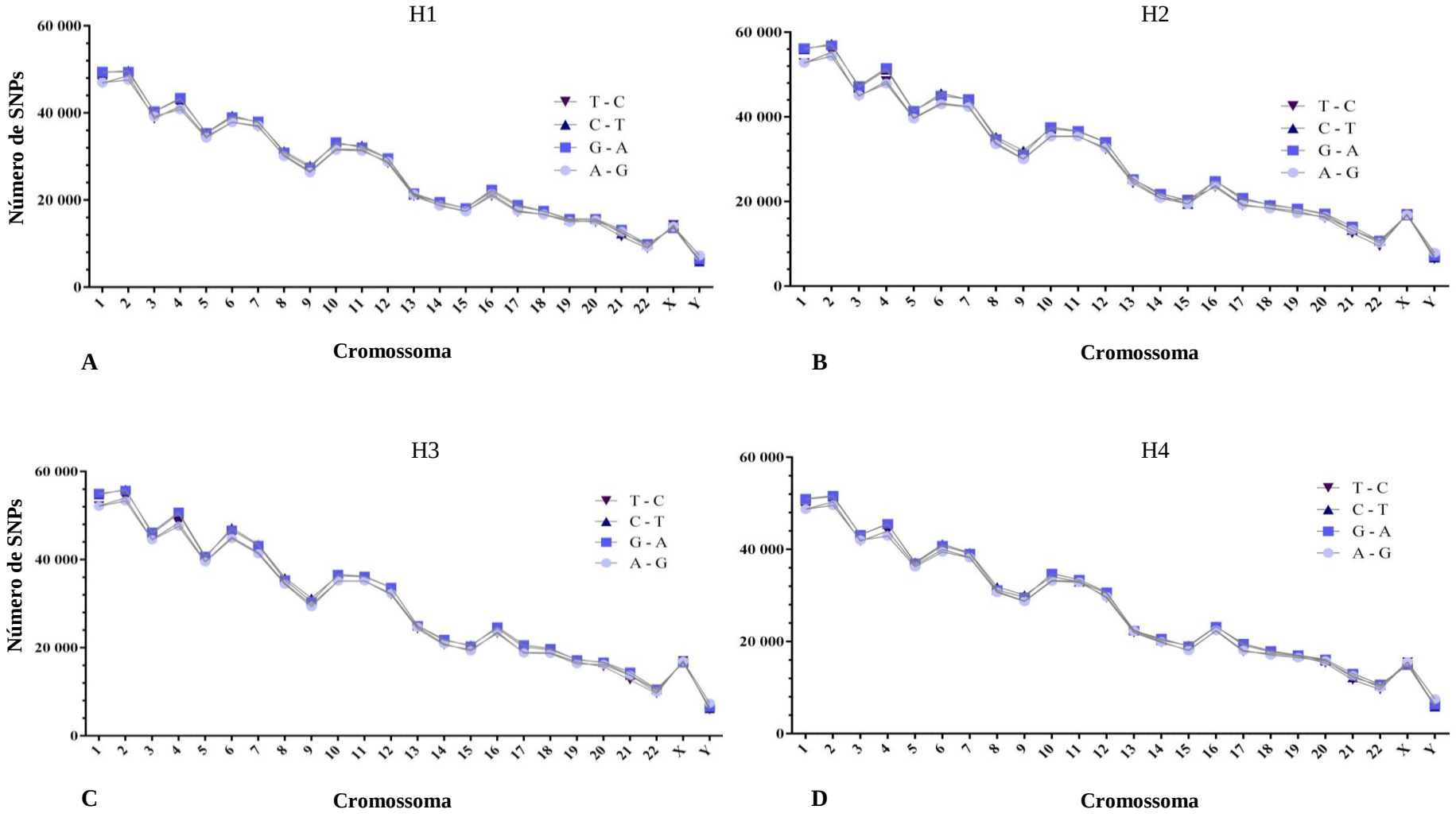


Figura 21: Distribuição dos tipos de Ts pelos cromossomas nucleares do genoma H1 (A), genoma H2 (B), genoma H3 (C) e genoma H4 (D).

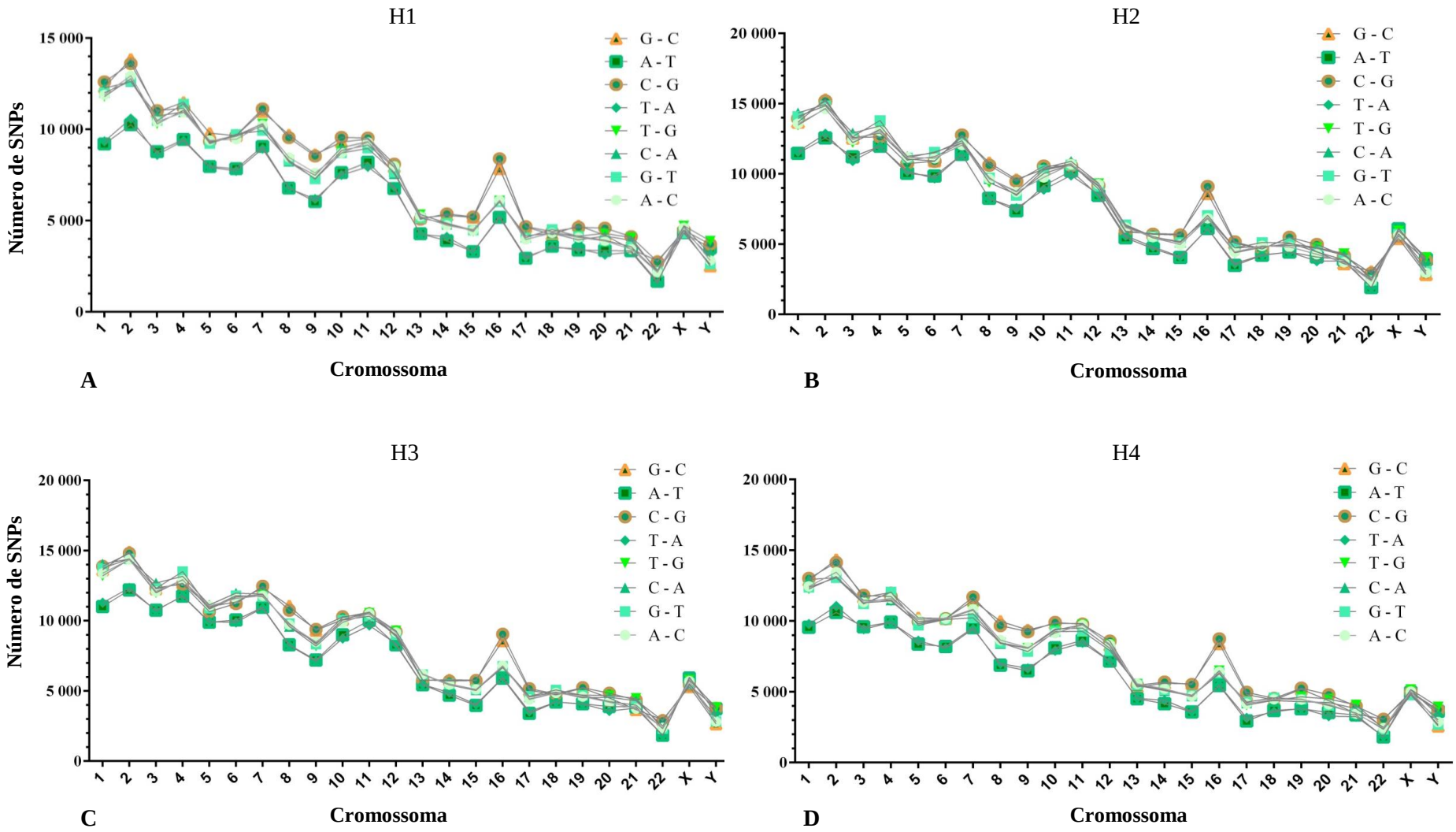


Figura 22: Distribuição dos tipos de Tv pelos cromossomas nucleares do genoma H1 (A), genoma H2 (B), genoma H3 (C) e genoma H4 (D).

2.4. Caracterização de INDELs

Após a caracterização dos SNPs, foram analisados os INDELs identificados no genoma H1 (353.768), os quais se distribuíram em 189.641 (53,6%) deleções e 164.127 (46,4%) inserções; no genoma H2 (536.250), os quais se distribuíram em 282.844 (52,7%) deleções e 253.406 (47,3%) inserções; no genoma H3 (503.832), os quais se distribuíram em 266.645 (52,9%) deleções e 237.187 (47,1%) inserções; e no genoma H4 (365.449), os quais se distribuíram em 195.700 (53,6%) deleções e 169.749 (46,4%) inserções. Analisando o comprimento de inserções e de deleções nestes genomas, verificou-se que existe um maior número de inserções e deleções de uma única base, observando-se respetivamente, no genoma H1 86.849 e 90.485; no genoma H2 144.184 e 143.308; no genoma H3 133.859 e 134.419; e no genoma H4 89.238 e 93.243 (Figuras 23, 24, 25 e 26). O comprimento médio dos INDELs nos quatro genomas foi de ≈2 pb para as inserções e ≈3 pb para as deleções. Efetuou-se o cálculo do rácio inserções/deleções e obteve-se o resultado de 0,87 (164.127/189.641) no genoma H1, 0,90 (253.406/282.844) no genoma H2, 0,89 (237.187/266.645) no genoma H3 e de 0,87 (169.749/195.700) no genoma H4.

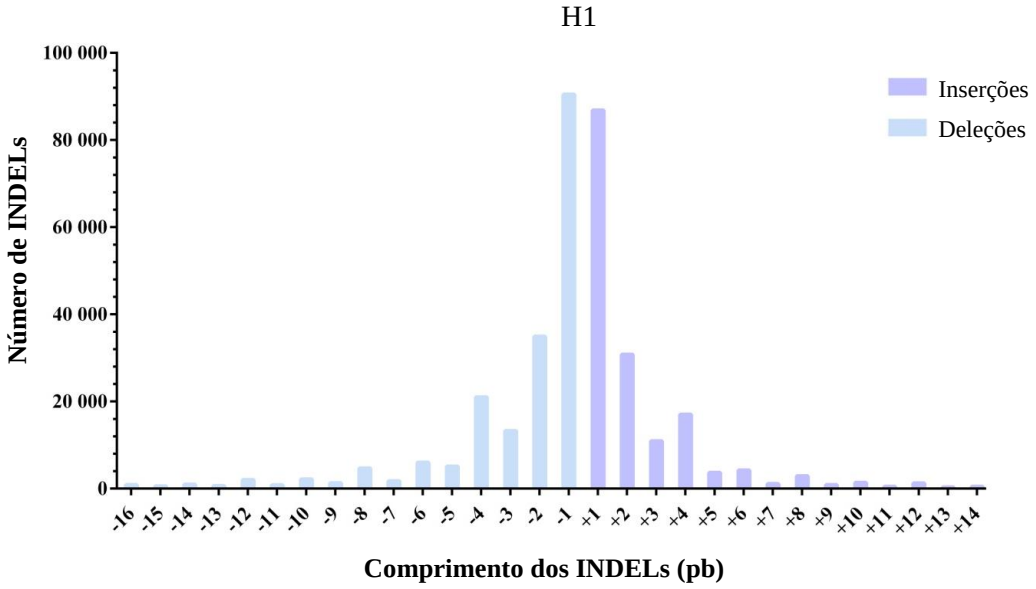


Figura 23: Caracterização do comprimento das inserções e deleções do genoma H1.

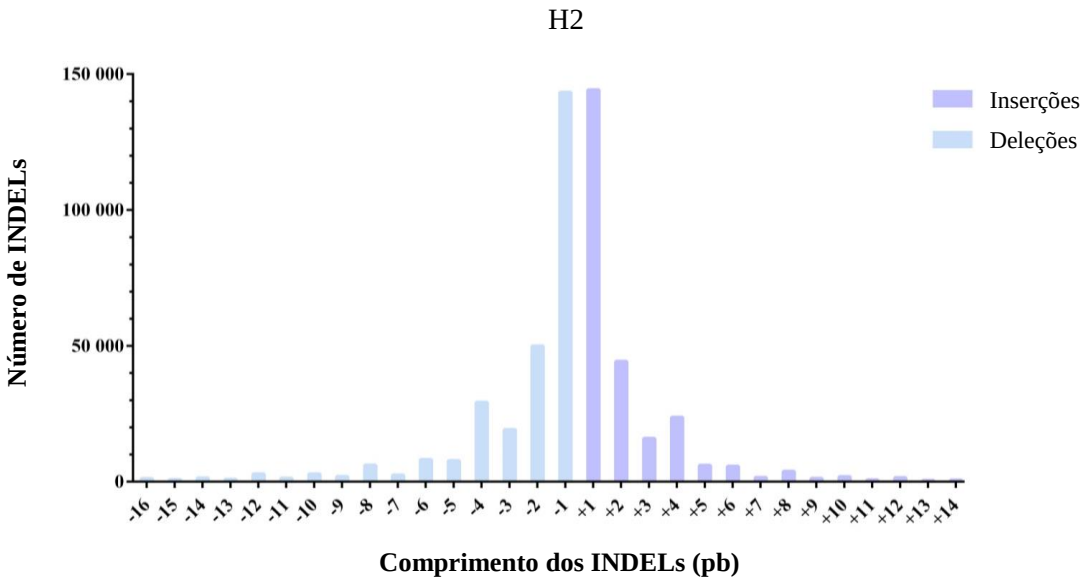


Figura 24: Caracterização do comprimento das inserções e deleções do genoma H2.

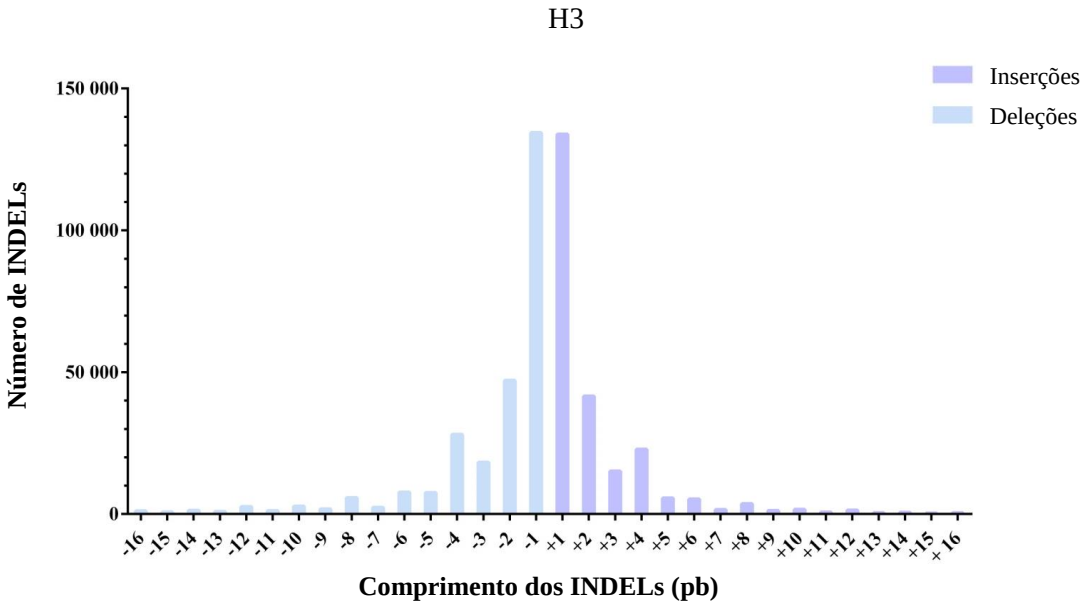


Figura 25: Caracterização do comprimento das inserções e deleções do genoma H3.

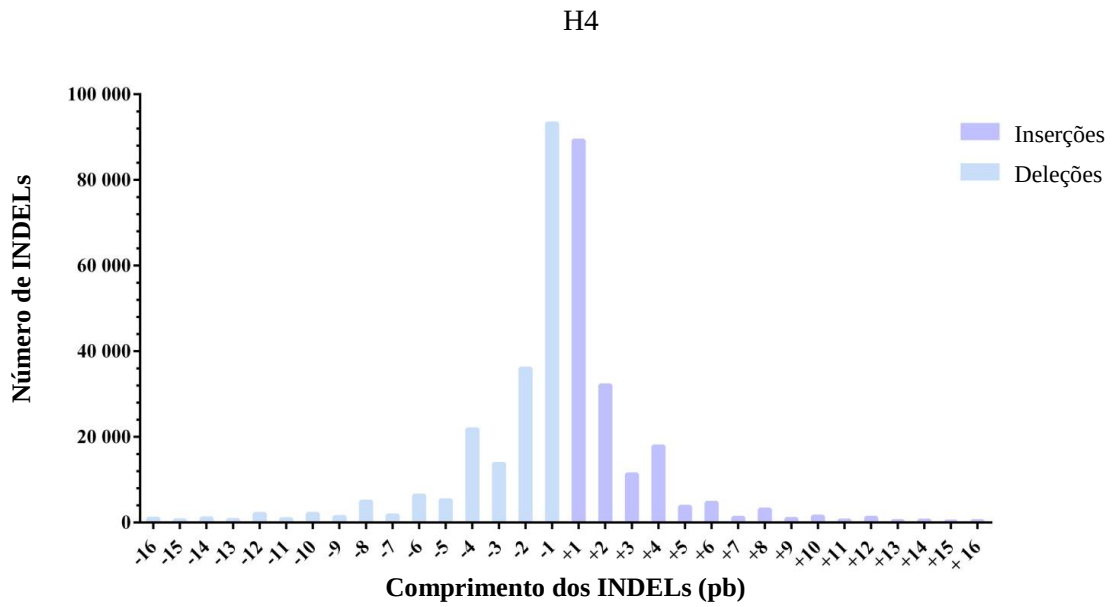


Figura 26: Caracterização do comprimento das inserções e deleções do genoma H4.

2.5. Estudo das variantes novas e conhecidas

Após a identificação das variantes conhecidas, identificadas com o respetivo rs pelo dbSNP137, observaram-se no genoma H1, H2, H3 e H4, 3.399.114 (88,8%), 3.913.212 (89,0%), 3.888.988 (89,4%) e 3.565.479 (88,9%) SNPs conhecidos, respetivamente, enquanto 427.784 (11,2%), 485.675 (11,0%), 461.884 (10,6%) e 445.297 (11,1%) SNPs dos mesmos genomas, respetivamente, foram classificados como variantes novas (Figura 27).

Considerando a distribuição dos SNPs novos e conhecidos nos quatro genomas estudados, calculou-se a razão entre o número de variantes por cromossoma e o número de pb que constitui cada cromossoma. Os resultados obtidos são apresentados na figura 28, na qual foi possível verificar que a taxa de variação de SNPs conhecidos é superior à de SNPs novos em todos os cromossomas, com exceção para o cromossoma Y, onde a tendência se inverte. É ainda possível verificar que o cromossoma Y apresentou a maior taxa de variação de SNPs novos em todos os genomas, com 0,05% no H1, H3 e H4, e uma taxa de 0,06% no H2. Os cromossomas 19 e 21 destacaram-se em todos os genomas analisados, apresentando a segunda maior taxa de variação de SNPs novos com 0,03%. A maior taxa de variação de SNPs conhecidos observou-se no genoma H1, nos cromossomas 10, 16, 20

e 21 com 0,13% cada; no genoma H2, nos cromossomas 4, 6, 16 e 19 com 0,15% cada; no genoma H3, nos cromossomas 4, 6, 16 e o 21 com 0,15% cada; e no genoma H4, nos cromossomas 16 e 19 com 0,14% cada (Figura 28).

No que respeita aos INDELs identificados com o respetivo rs pelo dbSNP137, observaram-se no genoma H1, H2, H3 e H4, 254.567 (72,0%), 363.134 (67,7%), 363.817 (72,2%) e 265.022 (72,5%) INDELs conhecidos, respetivamente, enquanto 99.197 (28,0%), 173.116 (32,3%), 140.015 (27,8%) e 100.427 (27,5%) INDELs, respetivamente, foram classificados como variações novas (Figura 27).

Atendendo à distribuição dos INDELs novos e conhecidos nos quatro genomas estudados, calculou-se a razão entre o número de variantes por cromossoma e o número de pb que constitui cada cromossoma. O rácio de INDELs conhecidos no genoma H1, H2, H3 e H4 foi superior em todos os cromossomas avaliados com exceção para o cromossoma Y, onde a tendência se inverte relativamente ao rácio de INDELs novos (Figura 29). Desta análise evidenciou-se o cromossoma 19 com a maior taxa de variação de INDELs novos, nos genomas H1, H2, H3 e H4, de 0,005%, 0,008%, 0,006%, e 0,005%, respetivamente, bem como com a maior taxa de variação de INDELs conhecidos, de 0,011%, 0,016%, 0,014% e 0,012%, respetivamente.

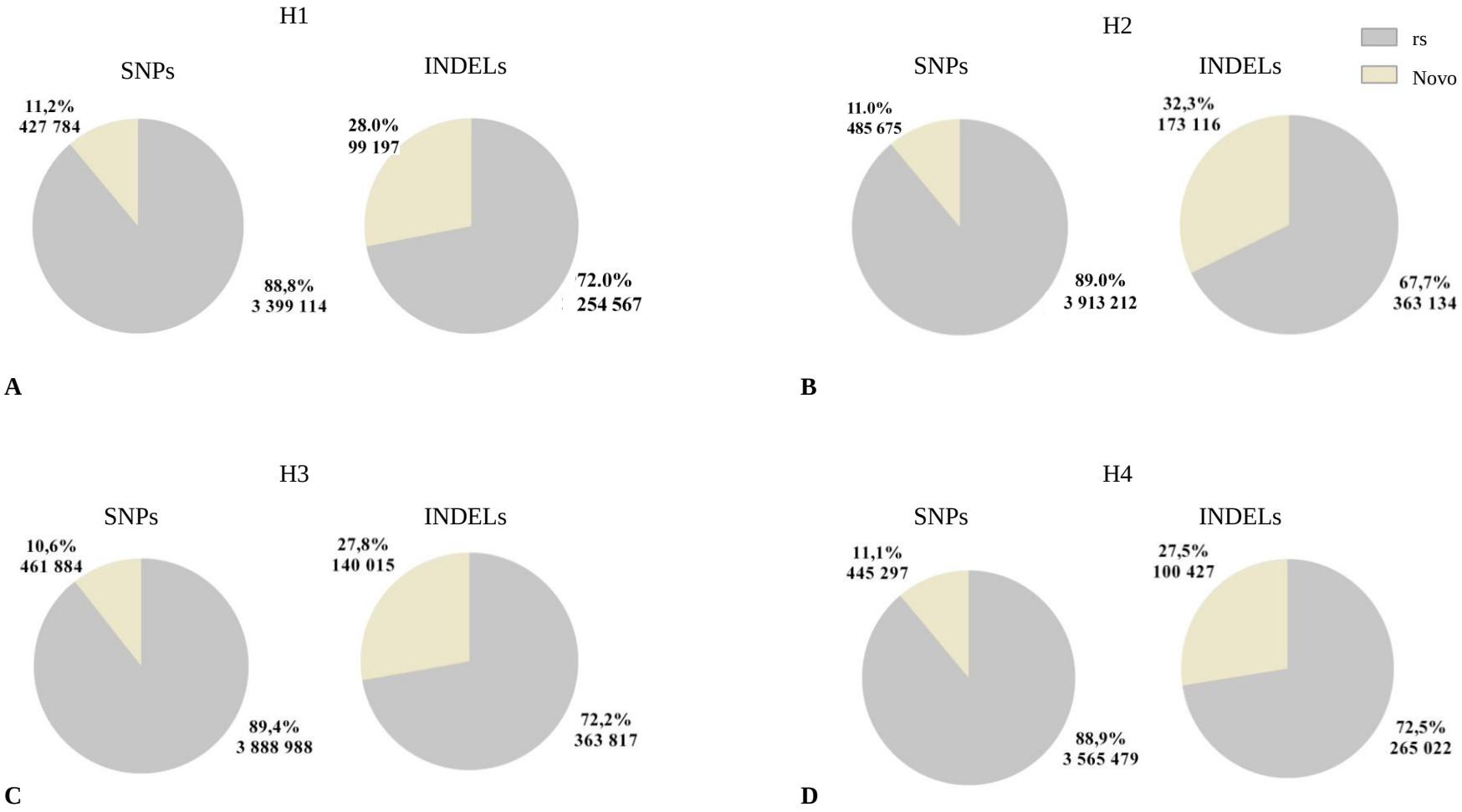


Figura 27: Análise de SNPs e INDELs novos e conhecidos no genoma H1 (A), H2 (B), H3 (C) e H4 (D).

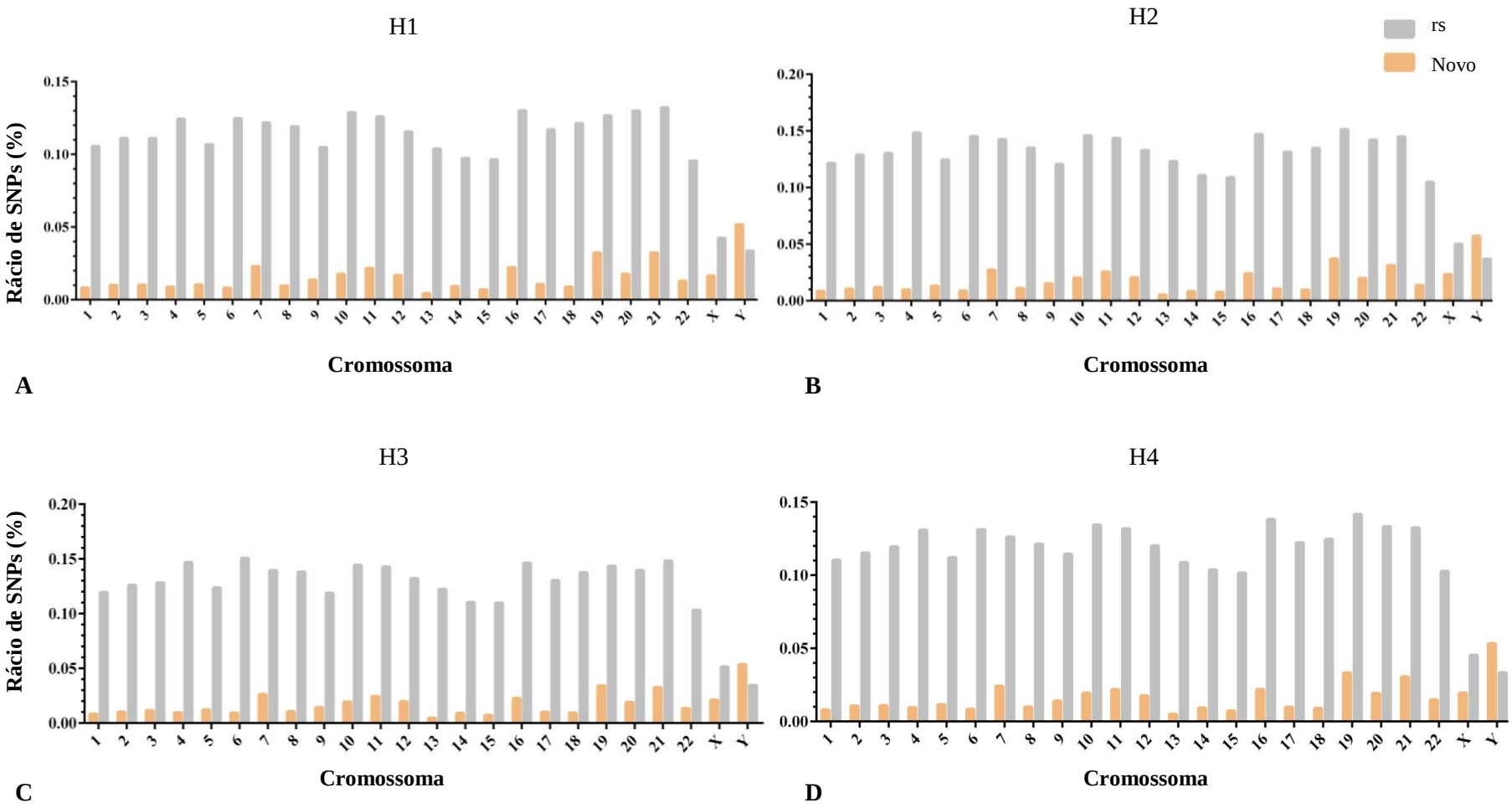


Figura 28: Rácio de SNPs novos e conhecidos nos diferentes cromossomas do genoma H1 (A), H2 (B), H3 (C) e H4 (D).

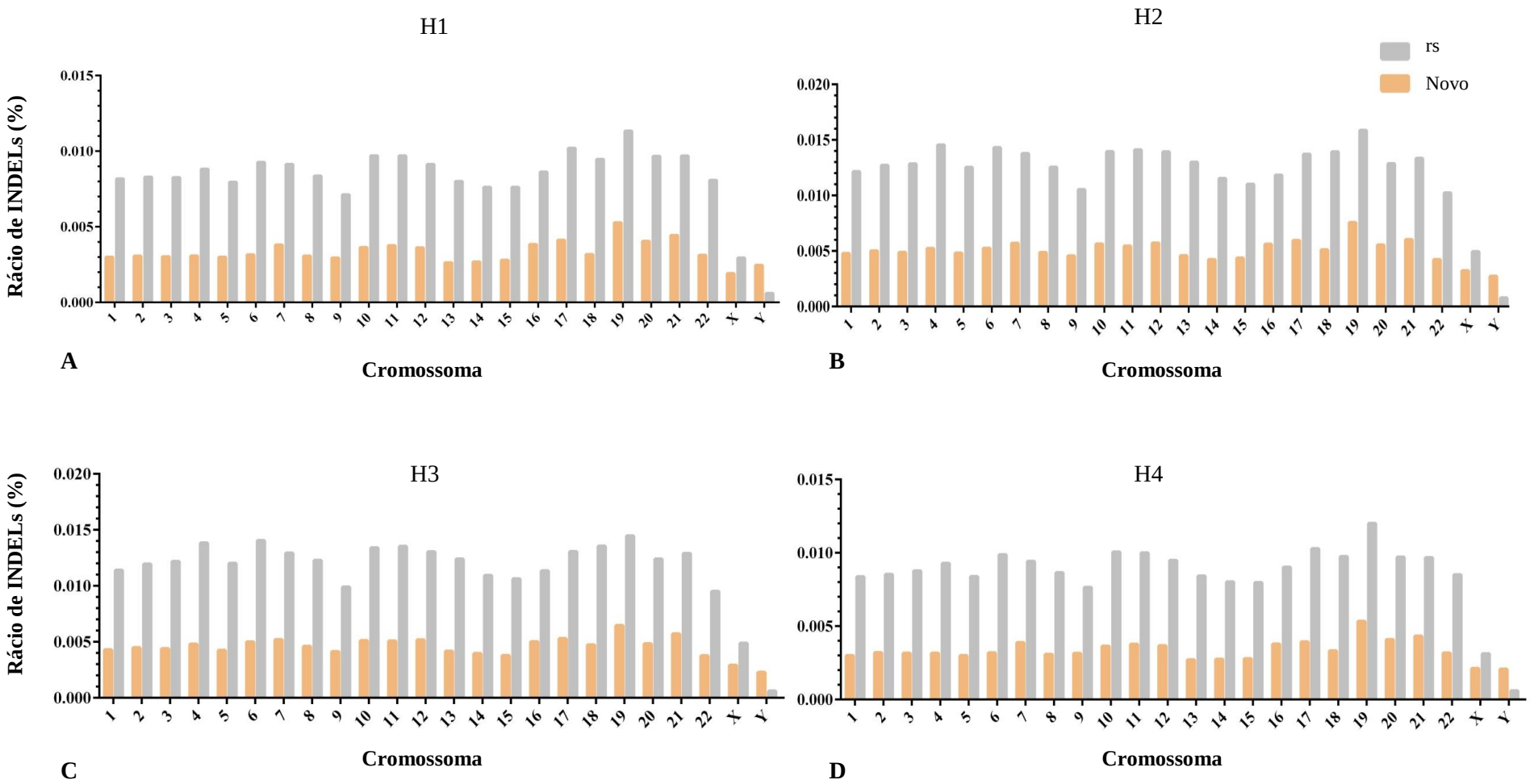


Figura 29: Rácio de INDELs novos e conhecidos nos diferentes cromossomas do genoma H1 (A), H2 (B), H3 (C) e H4 (D).

2.6. Estudo da zigotia das variantes

Em cada genoma foram caracterizadas as variantes pontuais quanto à zigotia, tendo sido identificadas as variantes em homozigotia (1/1 – quando os dois alelos são alterados e iguais entre si), e as variantes em heterozigotia (0/1 – quando um alelo está alterado e o outro é igual à referência; 1/2 – quando os dois alelos são alterados e diferentes estre si). Foram identificados, no genoma H1 1.384.779 SNPs 1/1, 2.438.717 SNPs 0/1 e 3.436 SNPs 1/2; no genoma H2 1.453.588 SNPs 1/1, 2.942.362 SNPs 0/1 e 2.974 SNPs 1/2; no genoma H3 1.472.964 SNPs 1/1, 2.874.811 SNPs 0/1 e 3.132 SNPs 1/2; e no genoma H4, 1.410.173 SNPs 1/1, 2.597.472 SNPs 0/1 e 3.173 SNPs 1/2. Relativamente aos INDELs identificaram-se, no genoma H1 157.425 INDELs 1/1, 189.198 INDELs 0/1 e 7.146 INDELs 1/2; no genoma H2 224.939 INDELs 1/1, 300.138 INDELs 0/1 e 11.174 INDELs 1/2; no genoma H3 219.930 INDELs 1/1, 274.245 INDELs 0/1 e 9.659 INDELs 1/2; e no genoma H4 165.499 INDELs 1/1, 191.867 INDELs 0/1 e 8.086 INDELs 1/2 (Figura 30). Destes valores decorre que as variantes pontuais em homozigotia nos genomas H1, H2, H3 e H4 corresponderam a 36,9%, 34,0%, 34,9% e 36,0% do total das variantes pontuais, respectivamente, enquanto as variantes em heterozigotia representaram 63,1%, 66,0%, 65,1% e 64,0%, respetivamente.

O rácio homozigotia/heterozigotia calculado foi de 0,58 (1.542.204/2.638.497) para o genoma H1, de 0,52 (1.678.527/3.256.648) para o genoma H2, de 0,54 (1.692.894/3.161.847) para o genoma H3, e de 0,56 (1.575.672/3.161.847) para o genoma H4.

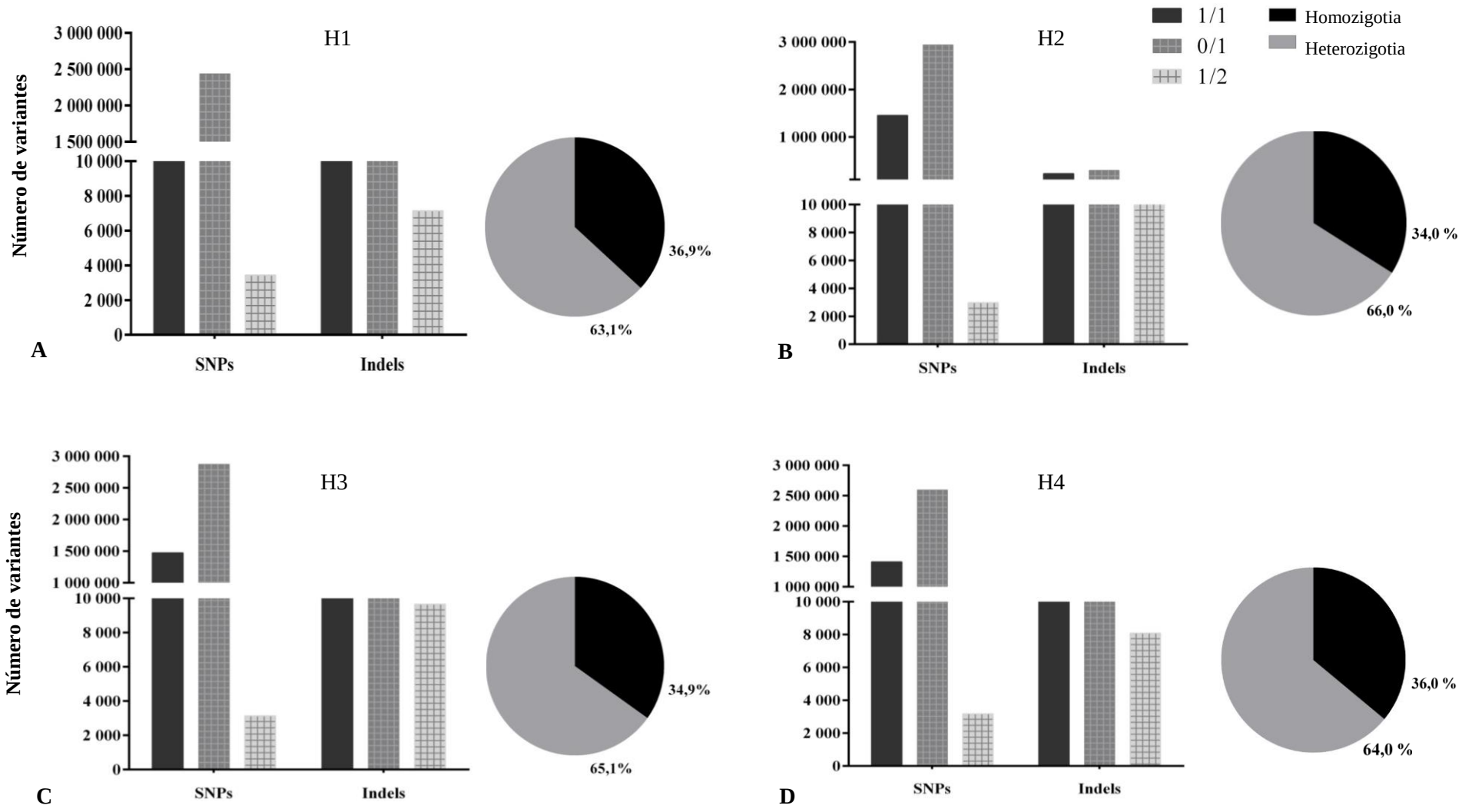


Figura 30: Anotação das variantes quanto à zigotia e frações homozigóticas e heterozigóticas no genoma H1 (A), H2 (B), H3 (C) e H4 (D).

3. Caracterização funcional das variantes exónicas

3.1. Anotação dos SNPs exónicos

Os SNPs presentes na região exónica foram estudados quanto ao seu efeito funcional, tendo sido considerados nos genomas H1, H2, H3 e H4, SNPs não sinónimos (*missense*) 11.662, 12.426, 12.459 e 12.010, respetivamente; SNPs que conduzem ao ganho de um codão *stop* (*nonsense*) 128, 128, 158 e 146, respetivamente; SNPs que conduzem à perda de um codão *stop* 49, 42, 49 e 52, respetivamente; e SNPs sinónimos 12.352, 12.604, 12.606 e 12.312, respetivamente (Tabelas 5 e 6).

O rácio *missense/nonsense* calculado foi de 91,11 (11.662/128) para o genoma H1, de 97,08 (12.426/128) para o genoma H2, de 78,85 (12.459/158) para o genoma H3, e de 82,26 (12.010/146) para o genoma H4.

Tabela 5: Anotação funcional dos SNPs identificados no genoma H1 e no genoma H2.

	H1				H2			
	Não sinónimo	**Ganho de *Stop***	**Perda de *Stop***	**Sinónimo**	**Não sinónimo**	**Ganho de *Stop***	**Perda de *Stop***	**Sinónimo**
Cromossoma 1	1.288	17	6	1.306	1.356	19	4	1.316
Cromossoma 2	670	7	3	762	752	9	3	796
Cromossoma 3	833	18	8	743	817	13	5	740
Cromossoma 4	425	5	4	459	473	2	5	496
Cromossoma 5	419	2	0	535	485	2	0	568
Cromossoma 6	847	6	3	736	926	10	3	769
Cromossoma 7	487	8	2	608	539	5	1	591
Cromossoma 8	328	4	1	416	362	3	1	397
Cromossoma 9	397	4	2	472	426	3	1	446
Cromossoma 10	411	2	1	465	516	3	2	505
Cromossoma 11	1.134	15	4	1.012	1.001	13	5	938
Cromossoma 12	612	2	3	622	605	6	2	651
Cromossoma 13	166	2	0	210	184	2	0	221
Cromossoma 14	356	1	0	393	358	1	0	407
Cromossoma 15	296	3	1	355	323	2	1	367
Cromossoma 16	399	2	2	467	498	1	2	555
Cromossoma 17	700	14	1	759	727	19	1	762
Cromossoma 18	164	0	0	184	160	0	0	170
Cromossoma 19	939	4	5	963	1.063	5	4	1.005
Cromossoma 20	209	1	1	284	238	0	0	292
Cromossoma 21	174	1	1	155	152	1	1	147
Cromossoma 22	223	4	0	274	259	4	0	263
Cromossoma X	182	6	1	172	202	5	1	198
Cromossoma Y	3	0	0	0	4	0	0	4
Total:	**11.662**	**128**	**49**	**12.352**	**12.426**	**128**	**42**	**12.604**

Tabela 6: Anotação funcional dos SNPs identificados no genoma H3 e no genoma H4.

	H3				H4			
	Não sinónimo	**Ganho de *Stop***	**Perda de *Stop***	**Sinónimo**	**Não sinónimo**	**Ganho de *Stop***	**Perda de *Stop***	**Sinónimo**
Cromossoma 1	1.339	18	5	1.264	1.280	14	6	1.231
Cromossoma 2	777	12	3	820	673	9	3	754
Cromossoma 3	833	19	8	747	865	22	8	804
Cromossoma 4	422	2	4	468	428	2	4	452
Cromossoma 5	451	2	0	545	405	3	0	524
Cromossoma 6	934	12	2	801	871	10	2	729
Cromossoma 7	577	3	2	570	505	4	2	552
Cromossoma 8	386	4	2	406	324	4	2	402
Cromossoma 9	460	4	2	467	448	2	1	468
Cromossoma 10	493	6	2	492	444	2	2	486
Cromossoma 11	1.049	19	6	935	1.020	14	5	875
Cromossoma 12	623	8	4	645	632	7	3	645
Cromossoma 13	161	1	0	225	174	2	0	186
Cromossoma 14	325	1	0	346	355	1	0	387
Cromossoma 15	349	6	0	408	307	6	1	398
Cromossoma 16	462	2	2	536	445	6	2	531
Cromossoma 17	745	18	1	764	766	16	1	754
Cromossoma 18	185	0	1	184	155	0	0	191
Cromossoma 19	1.078	8	2	1.045	1.095	12	7	1.018
Cromossoma 20	201	1	1	325	203	0	1	294
Cromossoma 21	151	2	1	149	171	1	1	144
Cromossoma 22	252	2	0	281	236	4	0	294
Cromossoma X	204	8	1	182	207	5	1	192
Cromossoma Y	2	0	0	1	1	0	0	1
Total:	**12.459**	**158**	**49**	**12.606**	**12.010**	**146**	**52**	**12.312**

3.2. Anotação dos INDELs exónicos

Os INDELs identificados na região exónica dos genomas estudados foram caracterizados quanto ao seu efeito funcional. O número de deleções e inserções *frameshift* encontrado foi, respetivamente, de 107 e 131 no genoma H1; de 124 e 147 no genoma H2; de 119 e 144 no genoma H3; e de 111 e 134 no genoma H4. O número de deleções e inserções *nonframeshift* determinado foi, respetivamente, de 134 e 89 no genoma H1; de 145 e 108, no genoma H2; de 154 e 112 no genoma H3; e de 136 e 87 no genoma H4. Foram também identificados 6, 4, 6 e 7 INDELs na região exónica dos genomas H1, H2, H3 e H4, respetivamente, que conduzem ao ganho de um codão *stop* (Tabelas 7 e 8). Destes valores decorre que os INDELs *frameshift* identificados nos genomas H1, H2, H3 e H4 corresponderam a 51,0%, 51,3%, 59,2% e 51,6% do total das INDELs, respectivamente, enquanto os INDELs *nonframeshift* representaram 47,8%, 47,9%, 49,7% e 46,9%, respetivamente. A restante fração de 1,2%, 0,8%, 1,1% e 1,5%, nos genomas H1, H2, H3 e H4, respetivamente, corresponde aos INDELs que conduzem ao ganho de um codão *stop* (Figura 31).

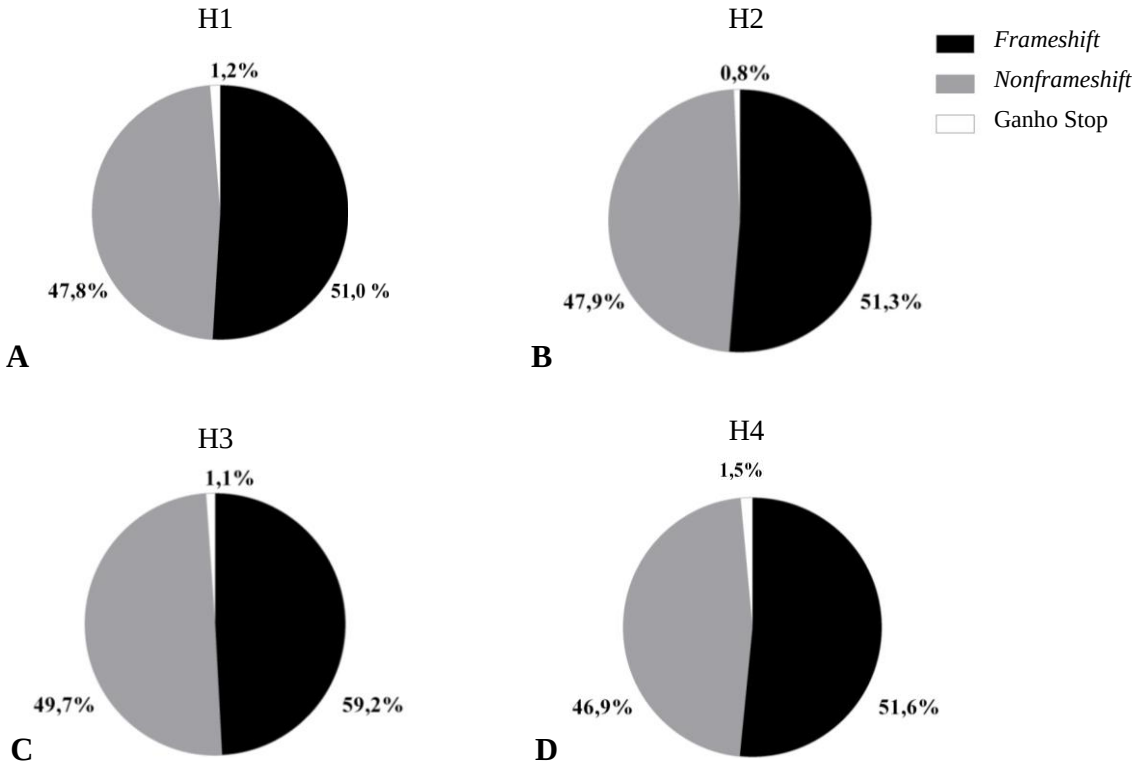


Figura 31: Anotação funcional dos INDELs identificados na região exónica do genoma H1 (A), do genoma H2 (B), do genoma H3 (C) e do genoma H4 (D).

Tabela 7: Anotação funcional dos INDELs identificados no genoma H1 e no genoma H2.

	H1					H2				
	Deleção *frameshift*	**Inserção *frameshift***	**Deleção *nonframeshift***	**Inserção *Nonframeshift***	**Ganho de *Stop***	**Deleção *frameshift***	**Inserção *frameshift***	**Deleção *nonframeshift***	**Inserção *Nonframeshift***	**Ganho de *Stop***
Cromossoma 1	12	6	12	10	1	8	5	13	11	1
Cromossoma 2	1	3	9	10	1	2	4	9	11	0
Cromossoma 3	7	6	8	8	3	7	8	7	6	1
Cromossoma 4	5	5	9	1	1	4	6	5	5	1
Cromossoma 5	3	6	9	1	0	1	6	9	1	0
Cromossoma 6	3	11	8	4	0	7	7	9	2	0
Cromossoma 7	6	11	6	4	0	6	11	10	7	0
Cromossoma 8	9	7	2	3	0	12	7	7	4	0
Cromossoma 9	0	1	3	5	0	4	1	9	3	0
Cromossoma 10	5	3	0	2	0	8	3	0	5	0
Cromossoma 11	11	18	11	4	0	11	19	9	5	0
Cromossoma 12	3	7	7	6	0	5	9	7	7	0
Cromossoma 13	1	3	2	0	0	1	4	1	0	0
Cromossoma 14	2	4	4	6	0	3	3	4	6	0
Cromossoma 15	3	0	4	1	0	3	3	5	0	0
Cromossoma 16	5	4	7	1	0	8	4	6	2	0
Cromossoma 17	11	15	11	6	0	11	17	5	9	0
Cromossoma 18	4	1	1	2	0	2	2	3	3	1
Cromossoma 19	7	11	8	6	0	10	14	12	11	0
Cromossoma 20	3	1	5	1	0	3	3	8	2	0
Cromossoma 21	4	1	2	3	0	4	1	0	3	0
Cromossoma 22	0	3	4	2	0	2	6	5	1	0
Cromossoma X	2	4	2	3	0	2	4	2	4	0
Cromossoma Y	0	0	0	0	0	0	0	0	0	0
Total:	**107**	**131**	**134**	**89**	**6**	**124**	**147**	**145**	**108**	**4**

Tabela 8: Anotação funcional dos INDELs identificados no genoma H3 e no genoma H4.

	H3					H3				
	Deleção *frameshift*	**Inserção *frameshift***	**Deleção *nonframeshift***	**Inserção *Nonframeshift***	**Ganho de *Stop***	**Deleção *frameshift***	**Inserção *frameshift***	**Deleção *nonframeshift***	**Inserção *Nonframeshift***	**Ganho de *Stop***
Cromossoma 1	9	6	11	12	1	10	5	14	7	1
Cromossoma 2	4	3	10	9	0	6	4	5	11	1
Cromossoma 3	10	6	11	7	2	9	8	9	7	3
Cromossoma 4	4	5	6	3	1	3	7	4	1	1
Cromossoma 5	3	6	10	4	0	3	5	9	4	0
Cromossoma 6	4	13	8	7	0	2	5	9	2	0
Cromossoma 7	8	10	7	4	1	7	10	6	6	1
Cromossoma 8	11	6	6	5	0	10	5	7	2	0
Cromossoma 9	2	4	9	3	0	5	2	8	3	0
Cromossoma 10	6	6	1	5	0	1	3	1	3	0
Cromossoma 11	12	19	8	5	0	12	15	9	3	0
Cromossoma 12	6	9	7	7	0	3	7	7	7	0
Cromossoma 13	1	0	0	0	0	0	1	1	0	0
Cromossoma 14	1	3	3	7	0	2	4	3	5	0
Cromossoma 15	2	2	7	2	0	2	1	4	3	0
Cromossoma 16	8	5	7	3	1	6	5	7	2	0
Cromossoma 17	9	17	6	7	0	9	17	7	5	0
Cromossoma 18	2	1	3	6	0	1	1	3	2	0
Cromossoma 19	7	12	15	8	0	9	15	8	8	0
Cromossoma 20	2	1	6	1	0	2	4	5	1	0
Cromossoma 21	5	1	2	2	0	4	1	1	1	0
Cromossoma 22	0	6	9	1	0	0	6	3	1	0
Cromossoma X	3	3	2	4	0	5	3	6	3	0
Cromossoma Y	0	0	0	0	0	0	0	0	0	0
Total:	**119**	**144**	**154**	**112**	**6**	**111**	**134**	**136**	**87**	**7**

3.3. Caracterização das consequências de variantes exónicas pelo SIFT, PolyPhen 2, LRT e *Mutation Taster*

As variantes exónicas foram analisadas por quatro ferramentas bioinformáticas, (SIFT, PolyPhen 2, LRT *e Mutation Taster*), e anotadas segundo o efeito previsto por cada ferramenta na função da proteína que resulta de cada gene alterado (Tabela 9).

Tabela 9: Anotação das variantes exónicas dos genomas H1, H2, H3 e H4, pelos algoritmos que preveem o efeito das variantes codificantes na função da proteína: SIFT (D – *Damaging*; T – *Tolerated*), PolyPhen 2 (P – *Probably damaging*; D – *Possibly damaging*; B – *Benign*), LRT (D – *Deleterious*; N - *Neutral*; U – *Unknown*) e Mutation Taster (A – *Disease causing automatic*; D - *Disease causing*; P – *Polymorphism automatic*; N – *Polymorphism*).

	SIFT		**PolyPhen 2**			**LRT**			***Mutation Taster***			
	D	**T**	**P**	**D**	**B**	**D**	**N**	**U**	**A**	**D**	**P**	**N**
H1	1.916	6.135	836	1.048	5.175	1.055	5.868	496	35	783	4.381	2.938
H2	2.119	6.534	957	1.142	5.390	1.156	6.252	538	37	853	4.686	3.155
H3	2.060	6.526	911	1.115	5.424	1.133	6.221	526	51	838	4.663	3.244
H4	1.942	6.285	853	1.110	5.153	1.057	5.986	522	40	730	4.536	3.015

As variantes exónicas com maior potencial deletério previsto pelos quatro *softwares* (D pelo SIFT; D e P pelo PolyPhen 2; D pelo LRT; e A e D pelo *Mutation Taster*) foram analisadas no sentido de identificar os genes nos quais se localizavam. Os genes identificados foram intersetados de modo a ser encontrado o conjunto que reunia as variantes categorizadas como deletérias pelos quatro *softwares* em simultâneo (Figura 32). Foram identificados 129 genes que continham 161 variantes previstas como deletérias, no genoma H1 e 116 genes em cada um dos genomas H2, H3 e H4, que continham 152, 175 e 151 variantes previstas como deletérias, respetivamente (Anexo 5).

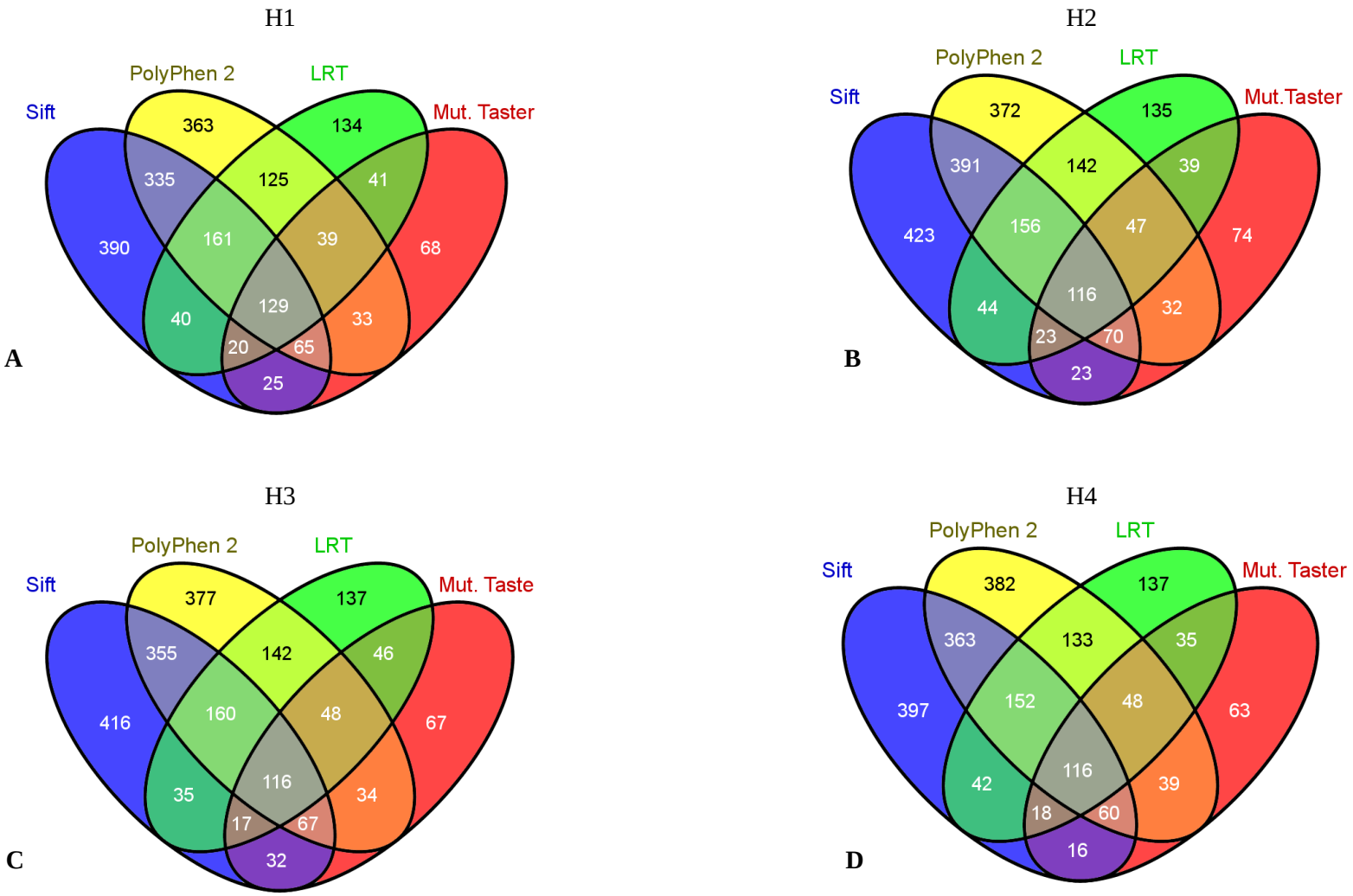


Figura 32: Diagrama de Venn dos genes identificados nos genomas H1 (A), H2 (B), H3 (C) e H4 (D), que continham as variantes com maior potencial deletério previsto pelos *softwares* SIFT, PolyPhen 2, LRT e *Mutation Taster*.

Os genes onde foram identificadas as variantes previstas como deletérias pelos quatro *softwares* em simultâneo, foram analisados de forma a identificar interações físicas e funcionais entre as proteínas codificadas por estes, através do programa STRING. Assim, obteve-se uma visão geral das proteínas que poderão estar, ou não, influenciadas pelas variantes identificadas, e que podem contribuir para o mesmo processo funcional.

No genoma H1 os 129 genes identificados foram analisados quanto às interações dos seus produtos proteicos (Anexo 7). Foram reportadas pelo STRING 36 interações, destacando-se três *clusters* com pelo menos três interações (Figura 33). Num dos *clusters* referidos encontravam-se as proteínas dos genes *SHC1*, *HDAC1* e *CTBP2*, que estão descritas como estando associadas à via da leucemia mieloide crónica, e também, em interação no mesmo *cluster*, as proteínas dos genes *LRRK2* e *HTRA2*, descritas como estando associadas à doença de Parkinson.

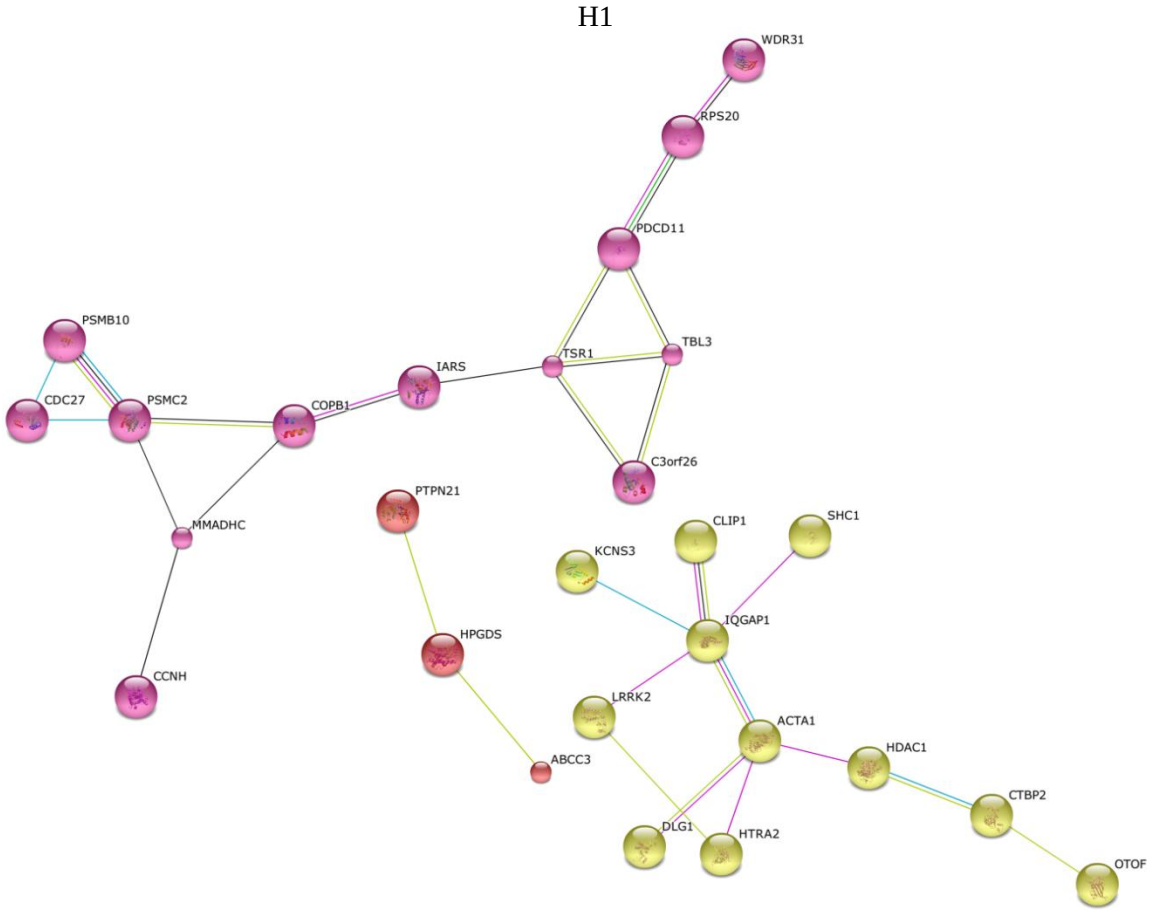


Figura 33: Análise das interações proteína-proteína referentes aos genes onde se localizam variantes consideradas deletérias no genoma H1 pelo STRING, onde se observam três clusters com mais de três interações.

No genoma H2, os 116 genes onde se localizavam as variantes previstas como deletérias pelos quatro *softwares* em simultâneo, foram analisados quanto às interações proteína-proteína dos seus produtos, tendo sido reportadas pelo STRING 25 interações (Anexo 8). Nesta análise evidenciou-se a existência de dois *clusters* de interação com mais de três interações cada (Figura 34). Num dos *clusters* supracitados, encontravam-se os genes *LAMA4* e *LAMA1*, estando as suas proteínas descritas como estando associadas ao cancro do pulmão. Foi ainda observada a interação entre as proteínas dos genes *VDR*, *ESR1* e *FOXM1* no outro *cluster* referido, sendo estas proteínas descritas com a mesma função molecular de regulação da transcrição do DNA.

H2

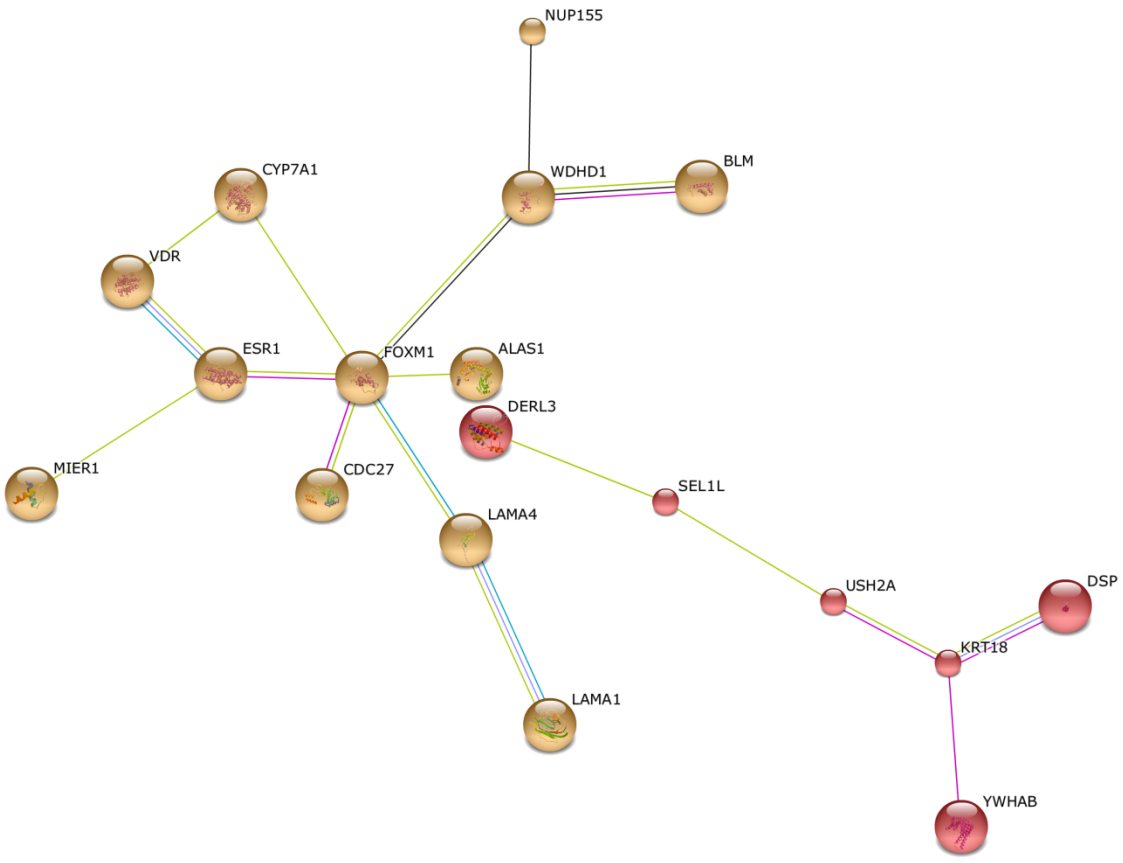


Figura 34: Análise das interações proteína-proteína referentes aos genes onde se localizam variantes consideradas deletérias no genoma H2 pelo STRING, onde se observam dois *clusters* com mais de três interações.

No genoma H3 os 116 genes, onde se localizam as variantes categorizadas como deletérias pelas quatro ferramentas bioinformáticas em simultâneo, foram analisados quanto às interações dos seus produtos proteicos, tendo sido reportadas pelo STRING 29 interações (Anexo 9). Nesta análise evidenciou-se a existência de dois *clusters* com mais de três interações (Figura 35). Num dos *clusters* supracitados, estavam presentes as proteínas dos genes *FOS* e *MAP2K3*, descritas como estando envolvidas em vias de sinalização via *toll-like receptors*. No mesmo *cluster* encontravam-se também as proteínas dos genes *NOTCH1*, *KAT2B* e *CTBP2*, também estas reportadas em associação a vias de sinalização. No segundo *cluster* referido, denota-se a interação descrita entre o produto dos genes *OR5P2*, *OR52J3* e *OR8B4*, todos eles associados à transdução do sinal olfativo.

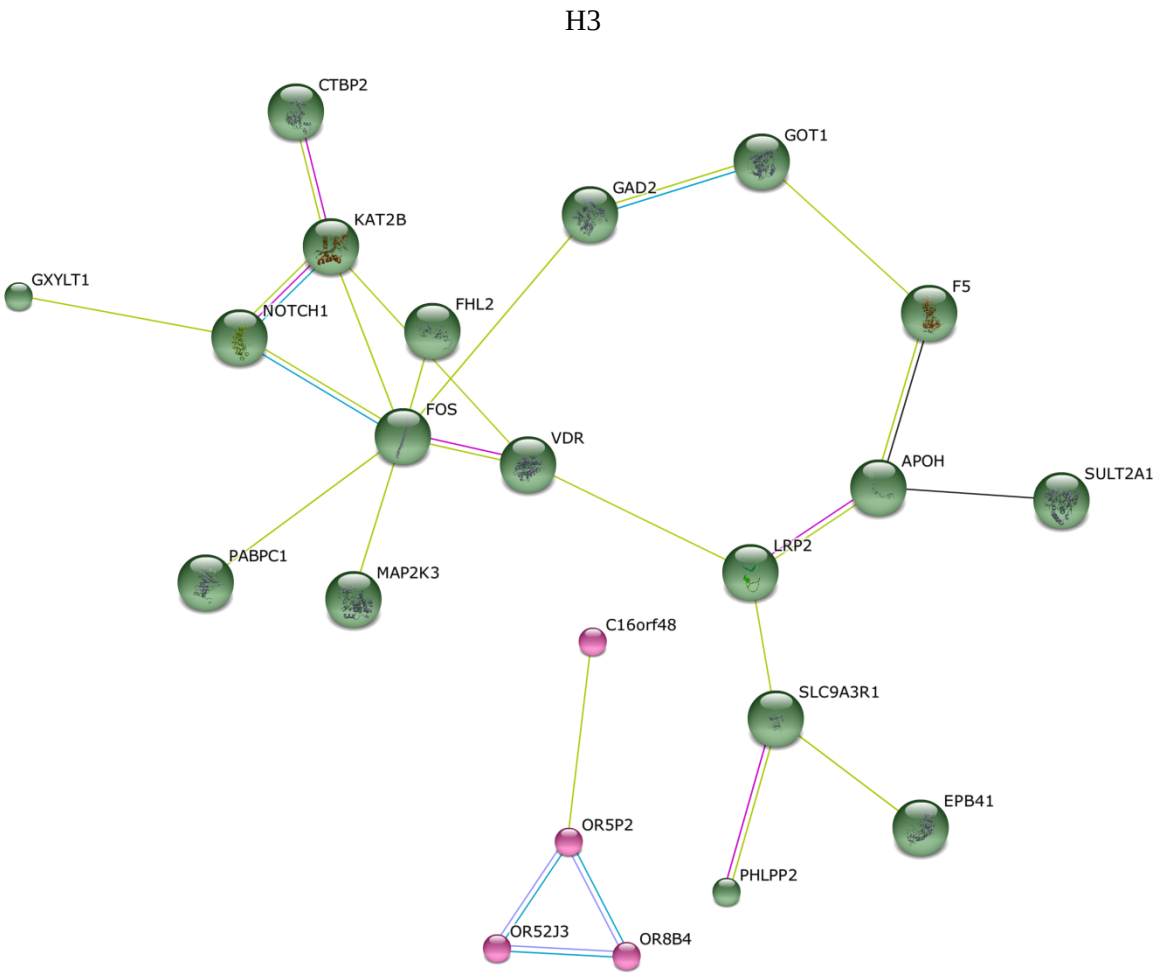


Figura 35: Análise das interações proteína-proteína referentes aos genes onde se localizam variantes consideradas deletérias no genoma H3 pelo STRING, onde se observam dois *clusters* com mais de três interações.

No genoma H4 os 116 genes, nos quais se identificaram as variantes categorizadas como deletérias pelos quatro *softwares* em simultâneo, foram analisados quanto às interações dos seus produtos proteicos, tendo sido reportadas pelo STRING 28 interações (Anexo 10). Nesta análise evidenciou-se a existência de três *clusters* com pelo menos três interações (Figura 36). Num dos *clusters* supracitados, encontravam-se referidas as proteínas dos genes *COL6A2, COL4A3, COL4A2* e *FN1*, as quais estão descritas como tendo funções ao nível da constituição da matriz extracelular. Os produtos dos genes *NOTCH1, COL4A3* e *COL4A2* estão também associados à função de regulação negativa da angiogénese e por isso destacou-se a sua interação reportada pelo STRING. As proteínas dos genes *COL4A2* e *FN1* destacam-se também pela sua associação descrita ao cancro do pulmão.

H4

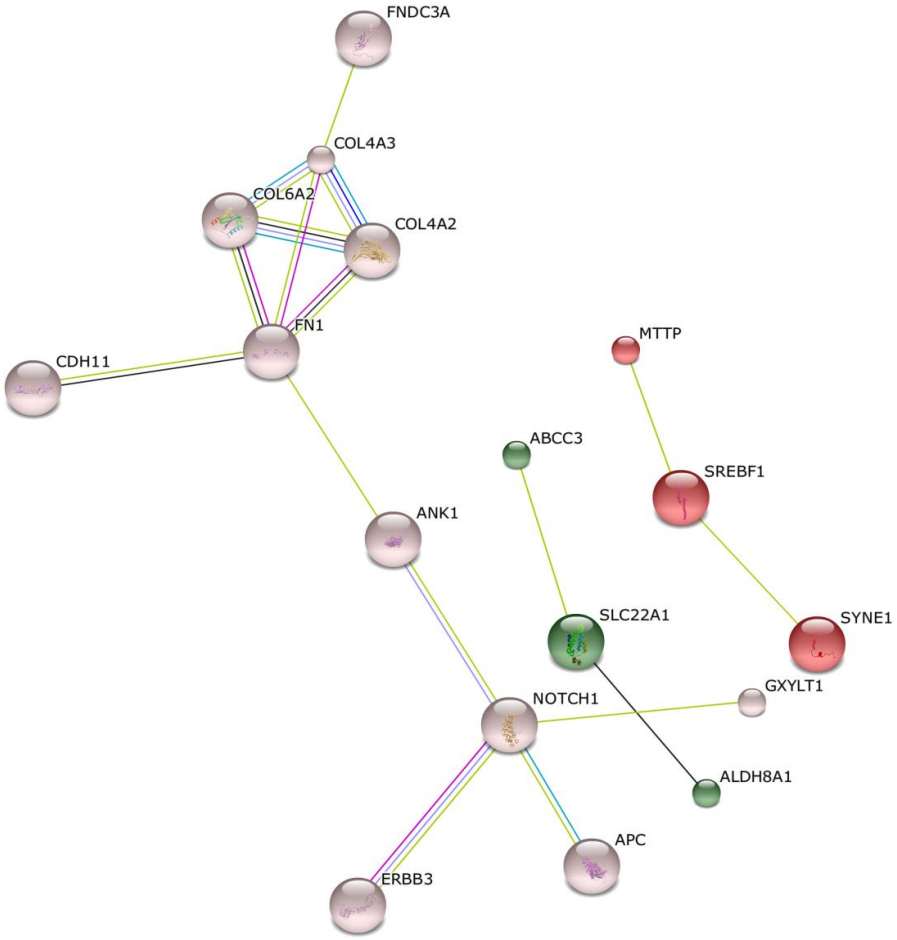


Figura 36: Análise das interações proteína-proteína referentes aos genes onde se localizam variantes consideradas deletérias no genoma H4 pelo STRING, onde se observam três *clusters* com pelo menos três interações.

3.3.1 Caracterização biológica e molecular dos genes onde se localizaram as alterações com maior potencial deletério previsto

Os genes nos quais se localizaram as variantes exónicas categorizadas como deletérias pelo SIFT, PolyPhen 2, LRT *e Mutation Taster*, foram caracterizados quanto à função biológica dos seus produtos proteicos, assim como quanto à função molecular, pelo PHANTER. A categorização dos genes do genoma H1, H2, H3, e H4 mediante a função biológica e função molecular das proteínas que destes resultam, encontra-se explanada na figura 37 e na figura 38, respetivamente.

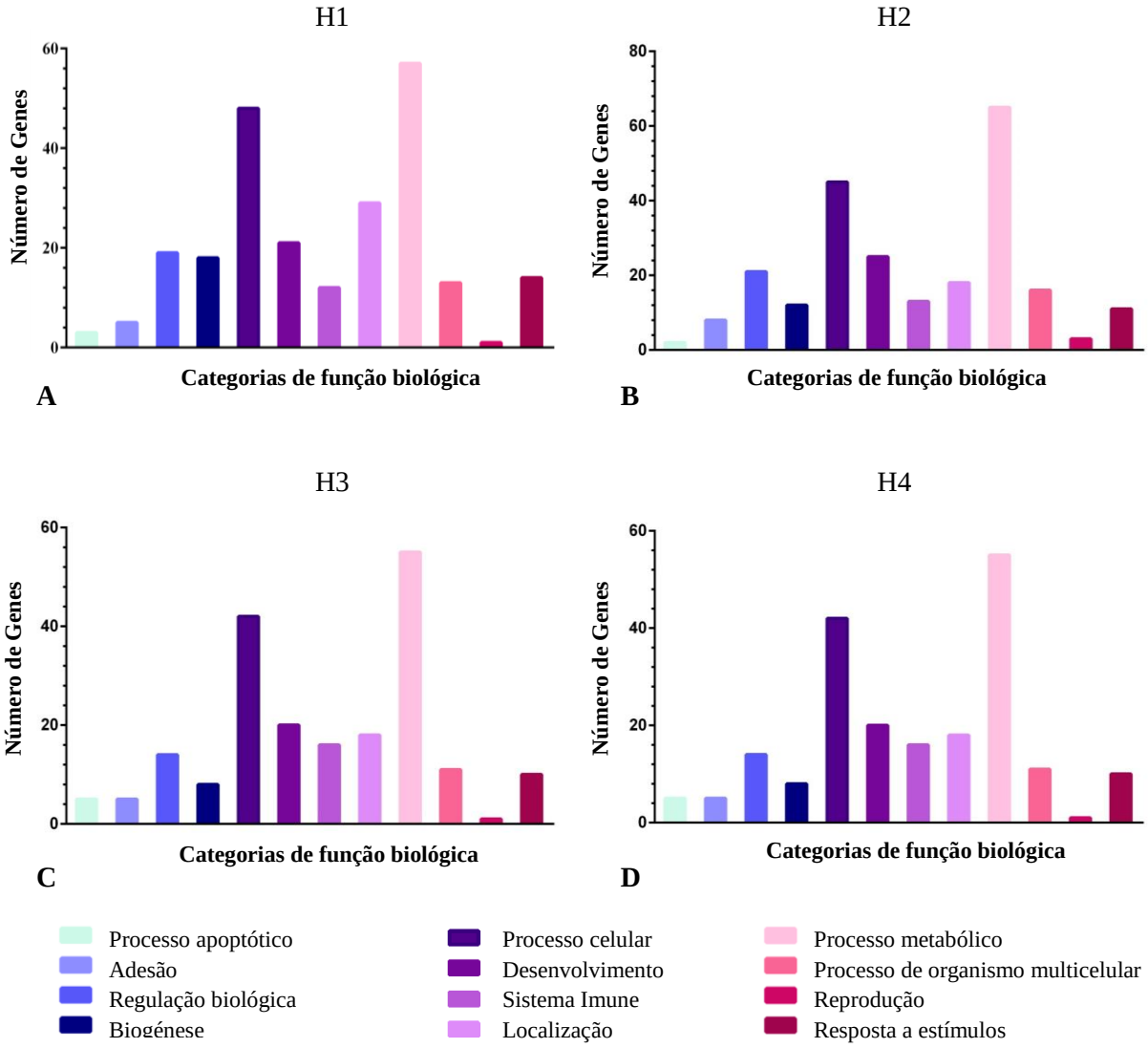


Figura 37: Caracterização da função biológica dos genes selecionados no genoma H1 (A), H2 (B), H3 (C) e H4 (D).

Nos genomas H1, H2, H3 e H4, destacou-se como principal função biológica das proteínas resultantes dos genes alterados, a categoria das funções ao nível do processo metabólico, seguindo-se as funções ao nível de processos celulares. Relativamente à função molecular destacou-se, igualmente nos quatro genomas analisados, que a atividade catalítica é a categoria com maior número de genes alterados pelas variantes identificadas nos genomas. No anexo 11, encontram-se especificadas as funções biológicas e as funções moleculares das proteínas resultantes de cada gene categorizado na figura 37 e 38.

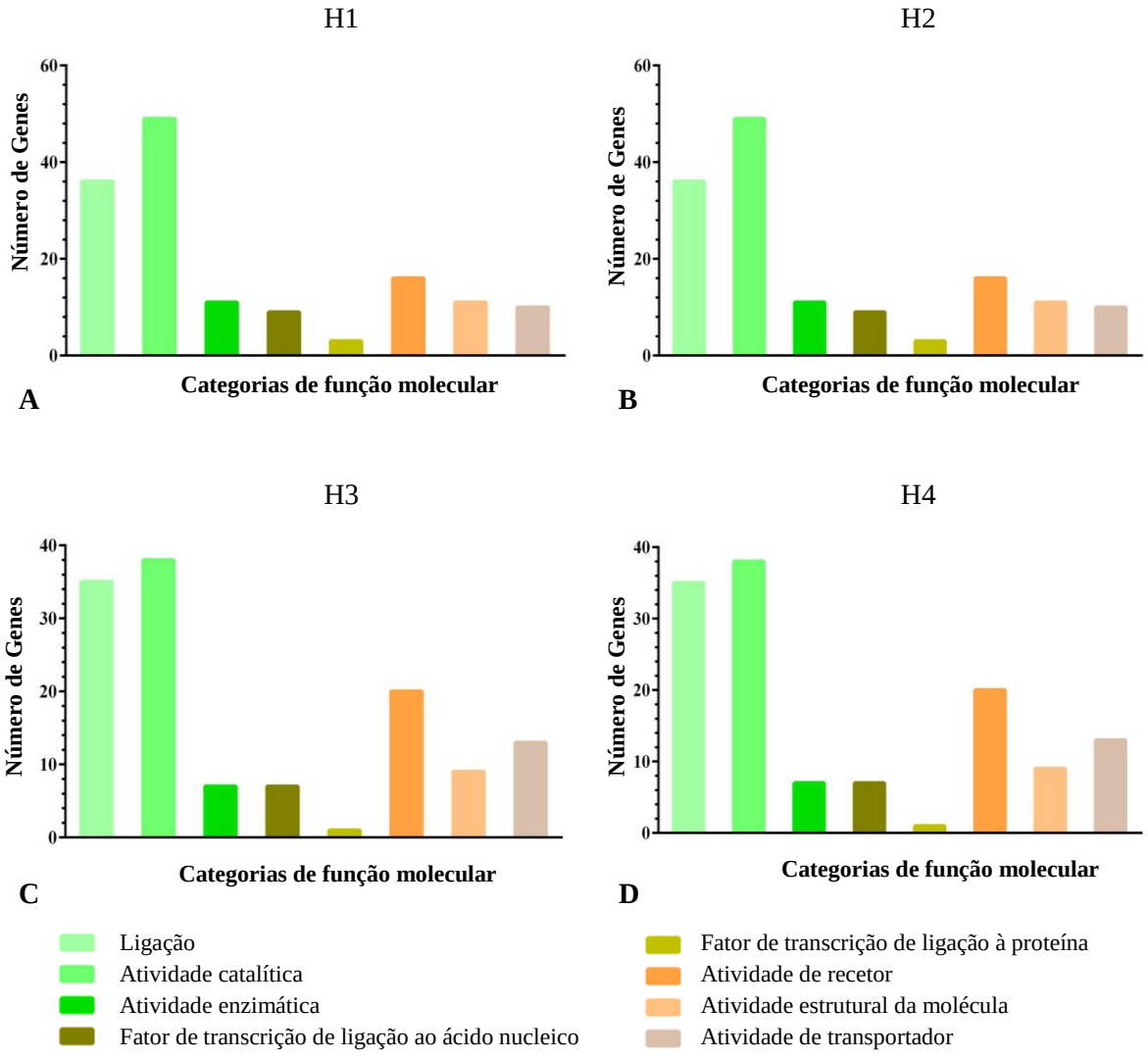


Figura 38: Caracterização da função molecular dos genes selecionados no genoma H1 (A), H2 (B), H3 (C) e H4 (D).

3.3.2. Posicionamento das alterações em zonas conservadas

As variantes exónicas categorizadas como deletérias pelo SIFT, PolyPhen 2, LRT *e Mutation Taster*, foram classificadas quanto à sua ocorrência em zonas conservadas e não conservadas do genoma humano, pelo Phylop. No genoma H1, 90,4% das variantes selecionadas para estudo funcional estavam em zonas conservadas do genoma e 9,6% estavam em zonas não conservadas. No genoma H2, 90,1% das variantes consideradas estavam em zonas conservadas enquanto 9,9% destas estavam em zonas não conservadas. No genoma H3, os valores mantêm-se semelhantes, 91,3 % das variantes foram identificadas em zonas conservadas e 8,7% em zonas não conservadas. No genoma H4, mantem-se novamente a tendência, com 88,2% das variantes selecionadas tendo sido localizadas em zonas conservadas e 11,8% em zonas não conservadas.

3.3.3. Caracterização das alterações quanto à frequência populacional

As variantes exónicas categorizadas como deletérias pelo SIFT, PolyPhen 2, LRT *e Mutation Taster*, foram analisadas contra as bases de dados dos projetos internacionais dos 1000 Genomas (população mundial) e ESP (populações americana e europeia), para determinar a sua frequência populacional.

A análise efetuada mostrou que no total das variantes exónicas previstas como deletérias nos genomas H1, H2, H3 e H4, uma fração de 54,4%, 40,7%, 50,0% e 44,1% respetivamente, foi identificada na população mundial do projeto dos 1000 Genomas, enquanto uma fração de 52,0%, 40,1%, 50,0% e 43,4% respetivamente, foi identificada na população americana/europeia do ESP. Das variantes dos genomas H1, H2, H3 e H4 encontradas nos 1000 Genomas, 92,4%, 92,4%, 97,3% e 91,7% destas, respetivamente, apresentaram uma MAF<5%, das quais 40,6%, 9,9%, 30,9% e 30,5% respetivamente apresentaram uma MAF<1%, enquanto 7,4%, 7,6%, 2,7% e 8,3% respetivamente, apresentaram uma MAF>5%. Relativamente às variantes dos genomas H1, H2, H3 e H4 encontradas na população americana/europeia do ESP, 96,9%, 96,9%, 97,3% e 98,3%, respetivamente, apresentaram uma MAF<5%, das quais 40,6%, 26,2%, 35,4% e 29,8% respetivamente apresentaram uma MAF<1%, enquanto 3,1%, 3,1%, 2,7% e 1,7% respetivamente, apresentaram uma MAF>5%.

3.3.4.Variantes associadas a fenótipo

As variantes exónicas de cada genoma foram analisadas de modo a determinar associações, previamente estabelecidas na bibliografia, entre estas variantes e fenótipos. No genoma H1 foi identificada, no gene *COL4A3* presente no cromossoma 2, uma substituição G/A (rs190598500) que está associada à síndrome de *Alport*. Na população mundial, esta variante tem uma MAF<1%. No gene *SLC45A2*, presente no cromossoma 5, foi identificada uma substituição C/G (rs16891982), que está relacionada com alterações ao nível da pigmentação dos olhos, pele e cabelos. Esta variante tem uma MAF de 44%, na população mundial. No gene *ZFYVE27* do cromossoma 10, foi identificada uma substituição G/T (rs35077384), que é associada à paraplegia espástica do tipo hereditário. Esta variante tem uma MAF de 2%. Também no cromossoma 10, mas no gene *HABP2*, foi identificada uma substituição G/A (rs7080536), associada a suscetibilidade para a estenose da carótida e a suscetibilidade para o tromboembolismo venoso. Esta variante tem uma MAF de 1%. No genoma H1 foi também identificada, no gene *LRRK2* do cromossoma 12, uma substituição C/T (rs33958906), que está associada à doença de Parkinson. Esta variante tem uma MAF de 2% na população mundial.

No genoma H2 foi identificada, no gene *BCHE* do cromossoma 3, uma substituição T/C (rs1799807) que é associada à deficiência de butirilcolinesterase e consequentemente à apneia pós-anestésica. Esta variante apresenta uma MAF de 1%. Tal como sucedeu no genoma H1, também no genoma H2 foi identificada a substituição C/G (rs16891982) do gene *SLC45A2*, que está relacionada com alterações ao nível da pigmentação dos olhos, pele e cabelos. No gene *DSP*, presente no cromossoma 6 do genoma H2, foi identificada uma substituição A/T (rs17604693) associada à cardiomiopatia. Esta variante tem uma MAF de 2%. Foi ainda identificada, no gene *PRF1* presente no cromossoma 10, uma variante que consiste na substituição G/A (rs35947132) e que está associada à linfo-histiocitose hemofagocítica familiar. A referida variante tem uma MAF de 2%.

No genoma H3, foi identificada a variação C/G (rs16891982) do gene *SLC45A2* relacionada com alterações ao nível da pigmentação dos olhos, pele e cabelos, tal como já havia sucedido no genoma H1 e H2. Foi ainda identificada no genoma H3, uma variante no

gene *TRAF3* presente no cromossoma 14, que consiste numa substituição C/T (rs143813189) e que está associada à suscetibilidade para encefalite por herpes simplex. A referida variante tem uma MAF<1%.

Da análise do genoma H4, resultou a identificação de uma variante no gene *COL4A3* presente no cromossoma 2, que consiste na substituição G/A (rs190598500), associada à síndrome de *Alport*, e que já havia sido identificada no genoma H1. No gene *AGXT* que se apresenta no cromossoma 2, foi identificada uma substituição C/T (rs34116584) que é associada à hiperoxalúria primária tipo 1. Esta variante tem uma MAF de 11%. Ainda neste genoma, foi identificada uma variante no gene *KRT75*, presente no cromossoma 12, que consiste na substituição C/T (rs2232398), associada à síndrome dos cabelos anágenos frouxos. A variante referida tem uma MAF<10%.

3.3.4.1. Farmacogenómica

Nos quatro genomas analisados, foi determinada a presença de variantes em genes específicos que conduzem à modulação da resposta terapêutica. No genoma H1 foi identificada uma variante, no gene *CMPK1* presente no cromossoma 1, que consiste na substituição G/T (rs35687416) e que está envolvida na resposta terapêutica à gemcitabina, associação esta descrita na população coreana. No genoma H3 foi identificada uma alteração no gene *XRCC1*, presente no cromossoma 19, que consiste na substituição G/A (rs1799782) e que está associada à resposta terapêutica à cisplatina e ciclofosfamida. No genoma H4, a alteração no gene *AGXT* do cromossoma 2, que consiste na substituição C/T (rs34116584), está envolvida na resposta terapêutica à fluorouracil, ao leucovosina e à oxoliplatina. Também no genoma H4, foi identificada uma alteração, no gene *SLC22A1* do cromossoma 6, que consiste na substituição G/A (rs34059508) e é associada à resposta terapêutica à metformina, ondansetrona e tropiretron. Uma outra variante está presente no gene *VDR* do cromossoma 12 e consiste na substituição A/G (rs2228570). Esta variante encontra-se associada a alterações na resposta terapêutica aos fármacos 1,25-dihidroxivitamina d3, calcipotriol, calcitriol e dexametasona, e foi identificada nos quatro genomas em análise. As associações anteriormente mencionadas foram descritas em estudos na população caucasiana. No genoma H2 não foi identificada nenhuma variante descrita como estando associada à modulação da resposta terapêutica.

4. Análise dos SNPs comuns aos quatro genomas

4.1. Distribuição dos SNPs por cromossoma

Após a análise dos quatro genomas individuais sequenciados, H1, H2, H3 e H4, foram caracterizados exclusivamente os SNPs que foram identificados em simultâneo nos quatro genomas. Da caracterização efetuada, foram identificados 1.697.587 SNPs comuns aos quatro genomas. O valor determinado distribuiu-se pelos cromossomas nucleares, tendo apresentado o cromossoma 2 o maior número de variantes partilhadas, com o valor de 130.339 SNPs, seguido pelo cromossoma 1 com um valor de 129.685 SNPs partilhados. O cromossoma 22 apresentou o menor número de SNPs partilhados pelos genomas, com o valor de 23.548 SNPs (Figura 39). A distribuição dos SNPs comuns aos quatro genomas pelos cromossomas nucleares mostra uma distribuição em proporção semelhante à distribuição da totalidade das variantes em cada genoma individual (Figuras 8, 9, 10, 11 e 39).

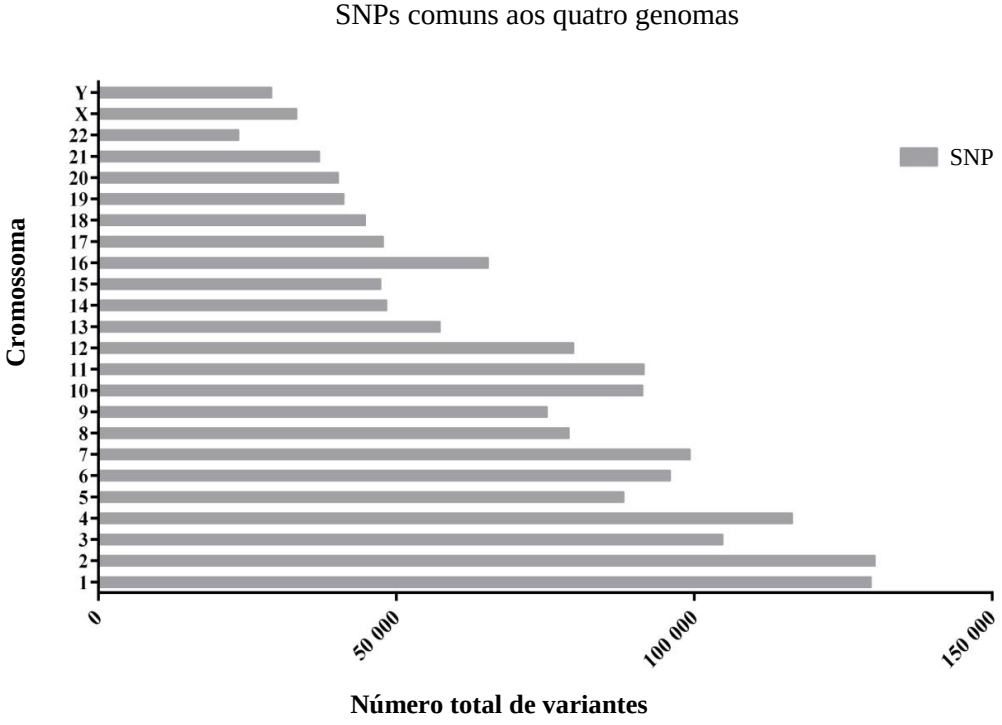


Figura 39: Distribuição dos SNPs pelos cromossomas nucleares, identificados nos quatro genomas (H1, H2, H3 e H4), em simultâneo.

Atendendo ao número de pb que compõem cada cromossoma, foi calculado o rácio de bases variantes relativamente ao número de SNPs comuns aos quatro genomas, por cromossoma (Tabela 10). Da análise resultou que o cromossoma 21 apresentou uma maior taxa de variação comum aos quatro genomas, com um rácio de 0,08%. Seguiu-se os cromossomas 10, 11, 16 e 19, cada um com um rácio de variação de 0,07%. O cromossoma X apresentou o menor rácio de variação, com 0,02%. A média do rácio de variação do genoma, relativamente à variação introduzida pelos SNPs partilhados pelos quatro genomas, foi de 0,05%.

Tabela 10: Distribuição de SNPs por cromossoma, comuns aos quatro genomas. Análise do rácio de bases variantes, relativamente ao número de SNPs em cada cromossoma nuclear.

	Comprimento (pb)	**SNPs**	**Rácio de bases variantes (%)**
Cromossoma 1	**249.250.621**	129.685	0,05
Cromossoma 2	**243.199.373**	130.339	0,05
Cromossoma 3	**198.022.430**	104.804	0,05
Cromossoma 4	**191.154.276**	116.488	0,06
Cromossoma 5	**180.915.260**	88.192	0,05
Cromossoma 6	**171.115.067**	95.998	0,06
Cromossoma 7	**159.138.663**	99.331	0,06
Cromossoma 8	**146.364.022**	78.988	0,05
Cromossoma 9	**141.213.431**	75.341	0,06
Cromossoma 10	**135.534.747**	91.365	0,07
Cromossoma 11	**135.006.516**	91.582	0,07
Cromossoma 12	**133.851.895**	79.773	0,06
Cromossoma 13	**115.169.878**	57.354	0,05
Cromossoma 14	**107.349.540**	48.407	0,05
Cromossoma 15	**102.531.392**	47.403	0,05
Cromossoma 16	**90.354.753**	65.434	0,07
Cromossoma 17	**81.195.210**	47.831	0,06
Cromossoma 18	**78.077.248**	44.788	0,06
Cromossoma 19	**59.128.983**	41.204	0,07
Cromossoma 20	**63.025.520**	40.280	0,06
Cromossoma 21	**48.129.895**	37.116	0,08
Cromossoma 22	**51.304.566**	23.548	0,05
Cromossoma X	**155.270.560**	33.272	0,02
Cromossoma Y	**59.373.566**	29.064	0,05
Total:	**3.095.677.412**	**1.697.587**	**0,05**

4.2. Distribuição dos SNPs comuns aos quatro genomas por região genómica

Os SNPs partilhados pelos quatro genomas em análise foram caracterizados relativamente à sua localização nas diferentes regiões genómicas (Figura 40) (Anexo 1). Concluiu-se que 1.096.460 (64,4%) dos SNPs estavam localizados na região intergénica e 601.127 (35,4%) dos SNPs estavam presentes na região intragénica. Dos 601.127 SNPs anotados na região intragénica, 10.428 (0,6%) estavam na região exónica e 503.159 (29,6%) estavam na região intrónica. Dos restantes SNPs presentes na região intragénica destacaram-se os SNPs presentes na região 5'UTR e na região 3'UTR, com 1.832 (0,1%) e 9.667 (0,6%), respetivamente; os SNPs presentes na região *upstream* e na região *downstream*, com 9.499 (0,6%) e 9.521 (0,6%), respetivamente; e os SNPs presentes na região ncRNA intrónica e na região ncRNA exónica, com 52.290 (3,1%) e 4.203 (0,2%), respetivamente.

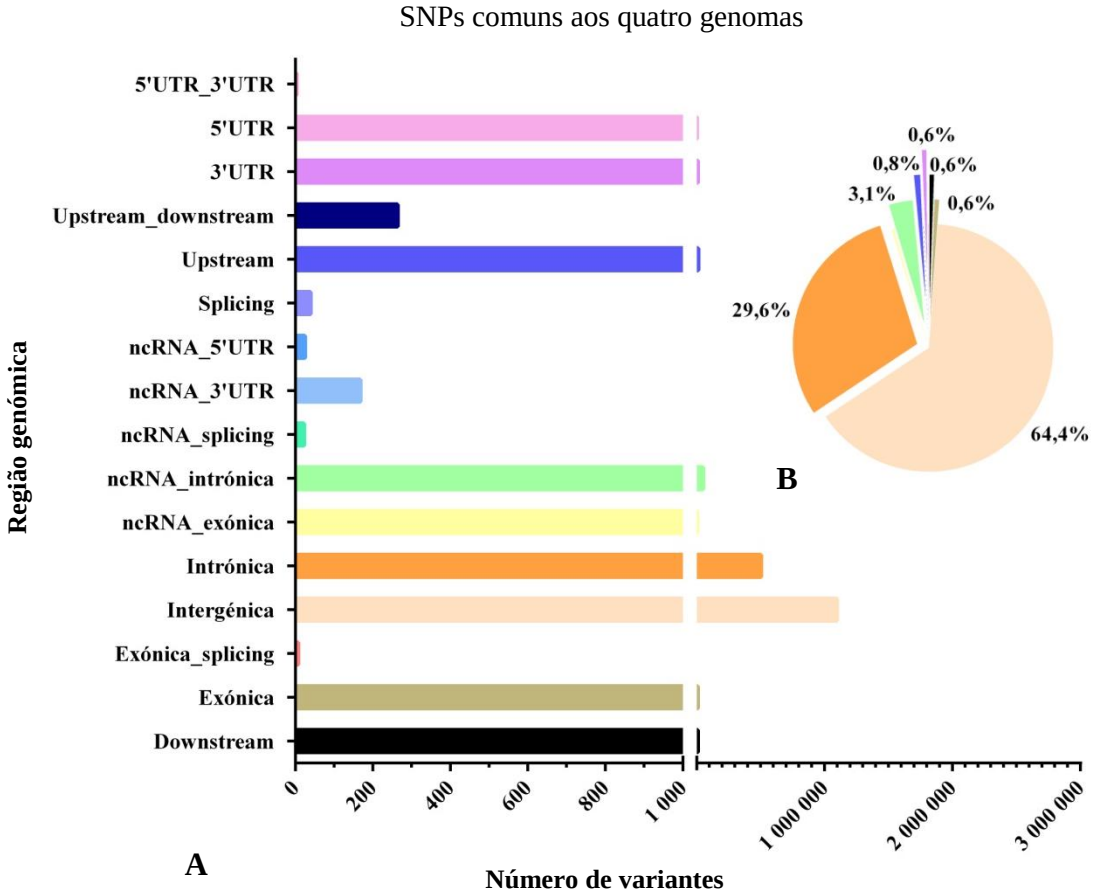


Figura 40: Distribuição dos SNPs comuns aos quatro genomas pelas diferentes regiões genómicas (A) e respetivos valores percentuais das regiões com maior abundância de SNPs (B).

4.3. Caracterização dos SNPs comuns aos quatro genomas

Os SNPs partilhados pelos quatro genomas sequenciados foram caracterizados relativamente ao tipo de SNPs, nomeadamente ao nível da ocorrência de Ts e Tv. Foram identificados 1.105.267 (65,1%) Ts e 592.320 (34,9%) Tv. Tal como sucedeu individualmente em cada genoma, verificou-se em todos os cromossomas um número superior de Ts comparativamente ao número de Tv, sendo que nos primeiros cromossomas essa diferença foi mais acentuada comparativamente aos restantes cromossomas (Figura 41). É possível observar a diferença supramencionada na figura 41, denotando-se que o cromossoma 1 apresentou 87.836 Ts e 41.849 Tv, enquanto o cromossoma Y apresentou 14.554 Ts e 14.510 Tv.

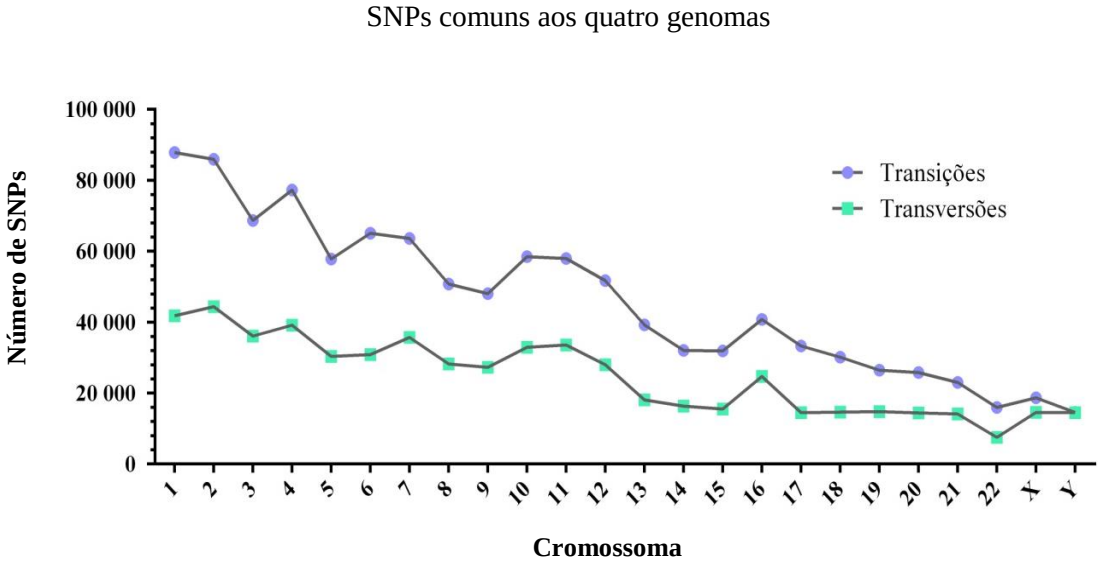


Figura 41: Distribuição de Ts e Tv comuns aos quatro genomas, pelos cromossomas nucleares.

Relativamente ao tipo de Ts, os SNPs partilhados pelos quatro genomas apresentaram uma ocorrência dos diferentes tipos na mesma proporção, ao nível de cada cromossoma, tal como havia sido verificado em cada genoma individual (Figura 42). Tendo sido efetuada a mesma análise para os diferentes tipos de Tv, verificou-se que a ocorrência destes sucede numa proporção semelhante, ao nível de cada cromossoma (Figura 43). Efetuou-se o cálculo do rácio Ts/Tv (1.105.267/592.320), tendo-se obtido o valor de 1,87.

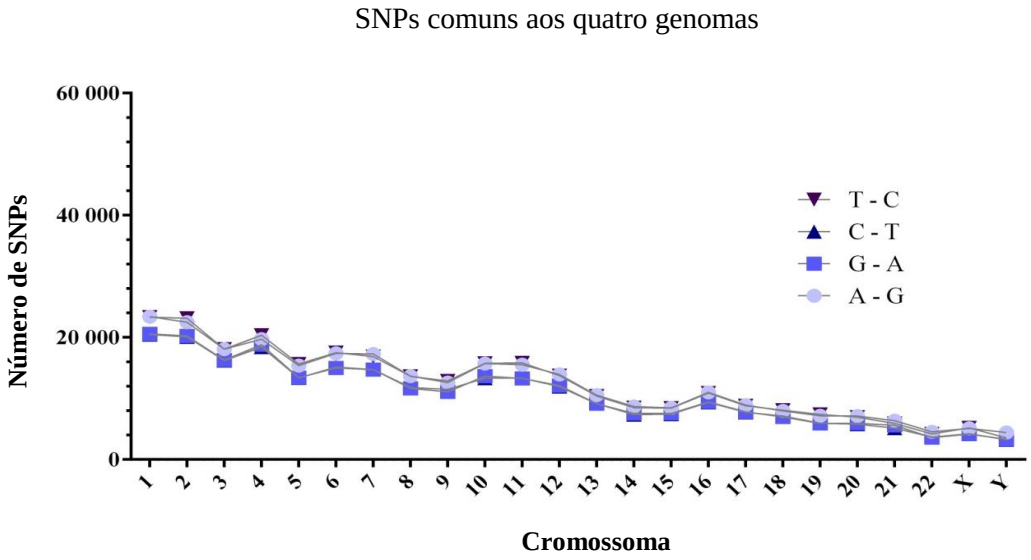


Figura 42: Distribuição dos tipos de Ts comuns aos quatro genomas, pelos cromossomas nucleares.

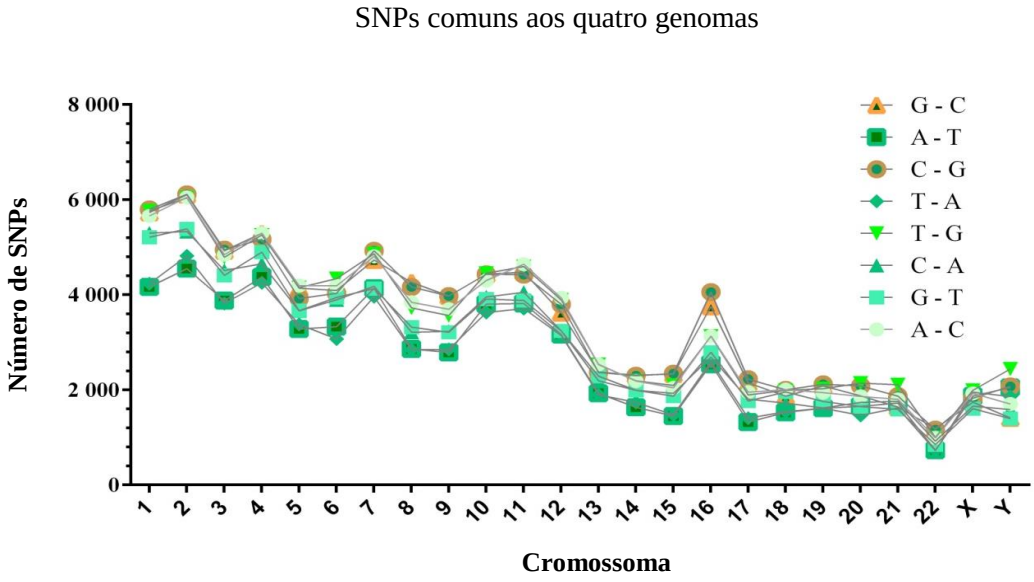


Figura 43: Distribuição dos tipos de Tv comuns aos quatro genomas, pelos cromossomas nucleares.

4.4. Estudo dos SNPs novos e conhecidos comuns aos quatro genomas

Após a identificação dos SNPs conhecidos, identificados na dbSNP137 com o respetivo rs, obteve-se o valor de 1.574.674 (92,6%) SNPs conhecidos, enquanto 125.913

(7,4%) dos SNPs partilhados pelos quatro genomas foram classificados como variantes novas, identificadas pela primeira vez (Figura 44). Considerando a distribuição dos SNPs novos e conhecidos, comuns aos quatro genomas, pelos cromossomas, calculou-se a razão entre o número de variantes por cromossoma e o número de pb que constitui cada cromossoma. Os resultados obtidos são apresentados na figura 45, na qual é possível verificar que o rácio de SNPs conhecidos foi superior ao rácio de SNPs novos em todos os cromossomas. Foi ainda possível verificar que o cromossoma Y apresentou o maior rácio de SNPs novos com 0,02%, seguido do cromossoma 19 e 21 com 0,01%. Os cromossomas 16 e 21 apresentaram o maior rácio de SNPs conhecidos, com 0,07% cada.

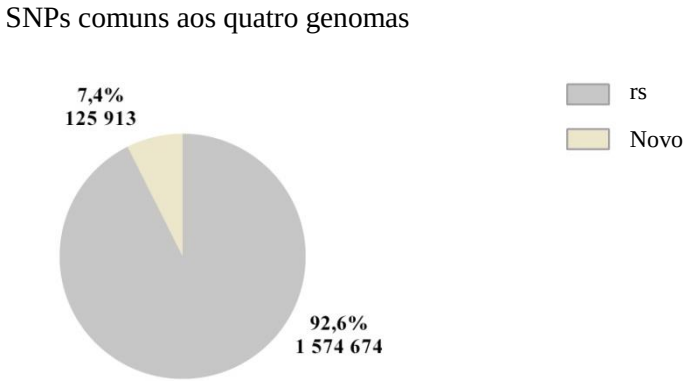


Figura 44: Análise de SNPs novos e conhecidos, partilhados pelos quatro genomas sequenciados.

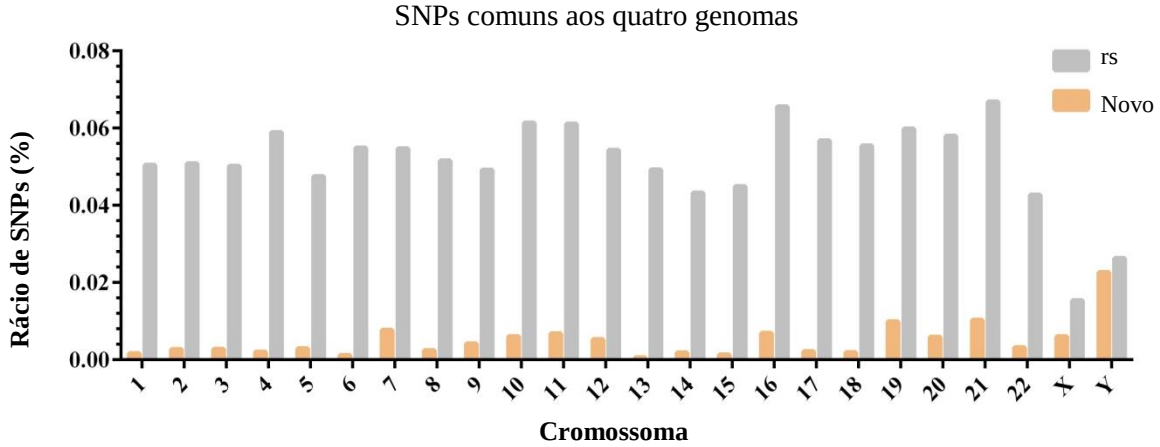


Figura 45: Rácio de SNPs novos e conhecidos, partilhados pelos quatro genomas, nos diferentes cromossomas nucleares.

4.5. Anotação dos SNPs exónicos comuns aos quatro genomas

Dos SNPs partilhados pelos quatro genomas sequenciados, presentes na região exónica, 5.004 foram SNPs não sinónimos (*missense*), 55 foram SNPs que conduziram ao ganho de um codão *stop* (*nonsense*), 9 SNPs levaram à perda de um codão *stop* e 5.126 foram SNPs sinónimos (Tabela 11). Efetuando o rácio *missense/nonsense* (5.004/55) obteve-se o valor de 90,98.

Tabela 11: Anotação funcional dos SNPs exónicos partilhados pelos quatro genomas.

	SNPs			
	Não sinónimo	**Ganho de *stop***	**Perda de *stop***	**Sinónimo**
Cromossoma 1	542	8	1	553
Cromossoma 2	285	7	0	301
Cromossoma 3	403	2	0	318
Cromossoma 4	165	0	2	186
Cromossoma 5	191	1	0	218
Cromossoma 6	336	3	0	315
Cromossoma 7	158	2	0	203
Cromossoma 8	141	3	0	151
Cromossoma 9	165	1	0	208
Cromossoma 10	200	2	0	205
Cromossoma 11	470	6	1	425
Cromossoma 12	300	2	2	285
Cromossoma 13	87	1	0	93
Cromossoma 14	134	1	0	139
Cromossoma 15	133	0	0	187
Cromossoma 16	166	0	1	207
Cromossoma 17	357	9	1	368
Cromossoma 18	62	0	0	74
Cromossoma 19	378	1	1	360
Cromossoma 20	69	0	0	111
Cromossoma 21	70	0	0	65
Cromossoma 22	97	2	0	95
Cromossoma X	94	4	0	59
Cromossoma Y	1	0	0	0
Total:	**5.004**	**55**	**9**	**5.126**

4.6. Caracterização de variantes exónicas comuns aos quatro genomas, pelo SIFT, PolyPhen *2*, LRT e *Mutation Taster*

As variantes exónicas partilhadas pelos quatro genomas foram analisadas por quatro ferramentas bioinformáticas, (SIFT, PolyPhen 2, LRT *e Mutation Taste*r), e anotadas segundo o efeito previsto por cada ferramenta na função da proteína que resulta

de cada gene alterado. Da anotação resultou: 616 SNPs anotados com D e 2.804 com T, pelo SIFT; 252 anotados com D, 219 com P e 2.235 com B, pelo PolyPhen 2; 322 variantes foram anotadas com D, 2.713 com N, 179 com U pelo *Mutation Taster*. As variantes exónicas comuns aos quatro genomas, com maior potencial deletério previsto pelos quatro *softwares* (D pelo SIFT; D e P pelo PolyPhen 2; D pelo LRT; e A e D pelo *Mutation Taster*), foram analisadas no sentido de identificar os genes nos quais se localizavam. Os genes identificados foram intersetados de modo a ser encontrado o conjunto que reunia as variantes comuns, categorizadas como deletérias pelos quatro *softwares* em simultâneo (Figura 46). Da interseção resultaram 11 genes que continham 28 variantes previstas como deletérias pelos quatro *softwares* em simultâneo (Anexo 6).

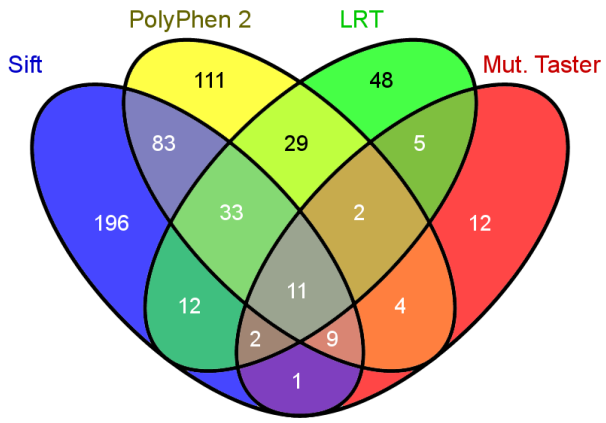


Figura 46: Diagrama de Venn dos genes que continham as variantes com maior potencial deletério previsto pelos *softwares* SIFT, PolyPhen 2, LRT e *Mutation Taster*, comuns aos quatro genomas

Os 11 genes que apresentaram as variantes com maior potencial deletério previsto pelos *softwares* SIFT, PolyPhen 2, LRT e *Mutation Taster*, comuns aos quatro genomas, foram analisados de forma a identificar interações físicas e/ou funcionais entre as proteínas codificadas por estes, descritas na base de dados IntAct e que foram representadas pelo programa Cytoscape. Assim, obteve-se uma visão geral das proteínas que poderão estar, ou não, influenciadas pelas alterações referidas e que podem contribuir para o mesmo processo funcional. Nesta análise foi reportada a interação indireta de 4 genes (CDC27, VDR, CLIP1 e PABPC1) dos 11 identificados (Figura 47). Na análise observou-se a interação da proteína CDC27, envolvida no ciclo celular, com a proteína RXRA, descrita

como estando envolvida na modulação da resposta à vitamina D e esta juntamente com a proteína SNW1, descrita como estando envolvida na transcrição da vitamina D, encontravam-se em interação com o recetor da vitamina D (VDR). A proteína SNW1 apresentou interação com a proteína THRAP3 e esta, por sua vez, com as proteínas CLIP1 e PABPC1 (Figura 47).

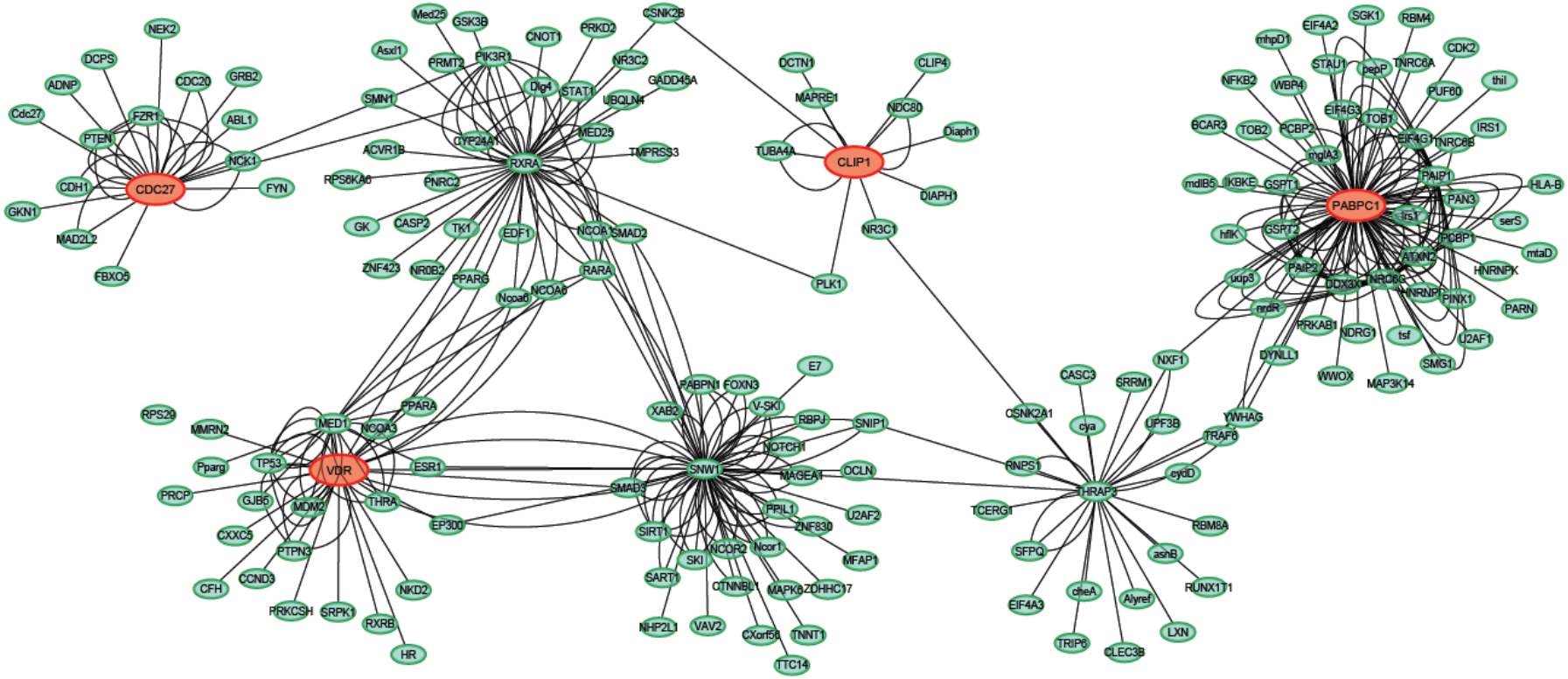


Figura 47: Análise das interações proteína-proteína, resultante do produto dos genes que contêm variantes previstas como deletérias, comuns aos quatro genomas, descritas na IntAct e representadas pelo Cytoscape. Observaram-se sete *clusters* em interação, sendo quatro desses *clusters* referentes a quatro proteínas cujos genes foram identificados como tendo as variantes previstas como deletérias, comuns aos quatro genomas (a vermelho).

4.6.1. Caracterização biológica e molecular dos genes onde se localizam as alterações comuns aos quatro genomas, com maior potencial deletério previsto

A categorização dos genes comuns aos quatro genomas, relativamente à função biológica e função molecular das proteínas que destes resultam, encontra-se explanada na figura 48. Destacou-se como principal função biológica das proteínas resultantes dos genes alterados, a categoria das funções ao nível do processo metabólico, seguindo-se as funções ao nível de processos celulares. Relativamente à função molecular destacou-se a atividade catalítica como categoria principal, com o maior número de genes alterados pelas variações identificadas nos quatro genomas. No anexo 11, encontram-se especificadas as funções biológicas e as funções moleculares das proteínas resultantes de cada gene categorizado na figura 48.

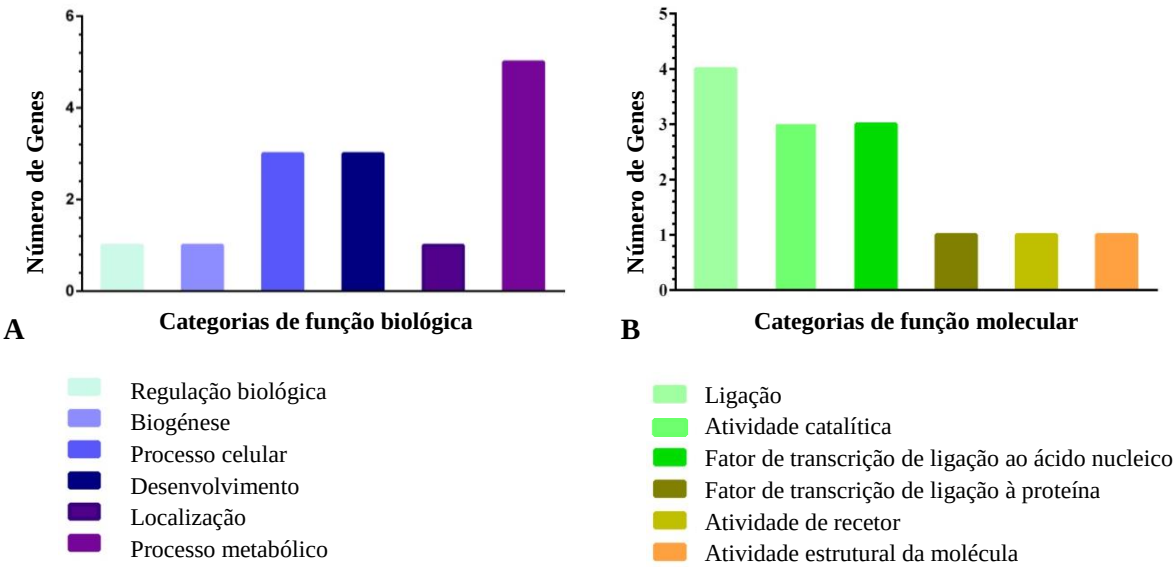


Figura 48: Caracterização da função biológica (A) e da função molecular (B) dos genes que contêm variantes comuns aos quatro genomas sequenciados.

4.6.2. Caracterização das alterações comuns aos quatro genomas, quanto à frequência populacional

As variantes exónicas comuns aos quatro indivíduos foram analisadas contra as bases de dados dos projetos internacionais dos 1000 Genomas e ESP. Das alterações analisadas, 32,1% das variantes exónicas selecionadas foram identificadas na população

mundial do projeto dos 1000 Genomas, sendo que destas 44,4% apresentaram uma MAF<5%, enquanto 55,6% apresentaram uma MAF>5%. As mesmas variantes exónicas foram caracterizadas quanto à presença na população americana/europeia do ESP, tendo 10,7% sido identificadas no ESP, das quais 66,7% apresentaram uma MAF<5% e 33,3% uma MAF>5%.

4.6.3. Variantes comuns aos quatro genomas associadas a fenótipo

As variantes exónicas comuns aos quatro genomas não apresentaram associações diretas a fenótipos de doença, descritos nas bases de dados *Ensembl* e OMIM. Denotou-se, contudo, a presença de uma variante no gene *VDR* do cromossoma 12, que conduz à modulação da resposta terapêutica. Esta variante, comum aos quatro genomas, consiste na substituição A/G (rs2228570) e encontra-se associada a alterações na resposta terapêutica aos fármacos 1,25-dihidroxivitamina d3, calcipotriol, calcitriol e dexametasona.

5. Caracterização de elementos reguladores não codificantes

5.1. rSNPs da região intrónica

A análise dos SNPs que têm impacto na regulação da expressão de genes foi considerada, nomeadamente em regiões não codificantes excluídas da análise de previsão do efeito que os SNPs têm na estrutura e consequente função da proteína alvo, anteriormente efetuada. Foram analisados os SNPs identificados na região intrónica dos genomas H1, H2, H3 e H4 de forma a identificar SNPs reguladores a nível transcricional e pós-transcricional, através da rSNPBase (Anexo 12). Em cada genoma, ≈ 80% dos SNPs identificados na região intrónica revelaram envolvimento em diversos tipos de regulação (Tabela 12). Foi verificado nos quatro genomas, que aproximadamente 20% dos rSNPs identificados em cada um foram atribuídos a um tipo de regulação transcricional proximal, isto é, a elementos reguladores que estão dependentes da sua proximidade genómica a locais de iniciação da transcrição para influenciar a própria transcrição. Nos quatro genomas, aproximadamente 31% dos rSNPs identificados referem associação a um tipo de regulação transcricional distal, isto é, a interações de cromatina em que diferentes locais de início da transcrição interagem mesmo estando distantes na sequência mas relativamente perto no espaço. Dos rSNPs identificados em cada genoma, 91% apresentaram também associação a regulação pós-transcricional mediada por proteína de ligação ao RNA, ou seja, que exercem regulação por intermédio da regulação de proteínas que se ligam a RNA para formar complexos ribonucleoproteícos essenciais à maturação do mRNA. Foi ainda observado que ≈72% de todos os SNPs identificados na região intrónica em cada genoma encontravam-se em LD (r^2 >0,8) com rSNPs.

Tabela 12: Identificação e caracterização do tipo de regulação de rSNPs na região intrónica dos genomas H1, H2, H3 e H4.

	SNPs intrónicos	**rSNPs na região intrónica**		**Tipos de regulação do rSNPs identificados**			
				SNPs em LD com rSNPs	Regulação Proximal	Regulação distal	Regulação mediada por uma proteína de ligação ao RNA
H1	**1.173.403**	909.008	77,5%	871.753	183.990	287.715	828.536
H2	**1.340.139**	1.077.436	80,4%	969.369	213.920	329.584	983.017
H3	**1.327.090**	1.066.537	80,4%	960.453	210.386	323.013	973.984
H4	**1.227.308**	979.697	79,8%	883.781	197.438	305.422	893.564

5.2. Variantes localizadas em miRNAs e o seu potencial efeito fenotípico

As variantes anotadas em miRNAs foram identificadas (Anexo 13), tendo sido encontrados nos genomas H1, H2, H3 e H4 um conjunto de 47, 60, 60 e 41 miRNAs alterados, respetivamente. Do conjunto de miRNAs alterados em cada genoma, foram selecionados os miRNAs descritos em simultâneo nas bases de dados miRTarBase versão 4.5 e Microcosm versão 5, de forma a serem identificados os respetivos genes alvos de regulação (Figuras 49, 51, 53 e 55).

As proteínas resultantes dos genes alvo da regulação dos miRNAs que continham variantes identificadas em cada genoma, foram integradas nas redes de interação proteína-proteína resultantes do produto dos genes codificantes que continham variantes previstas como deletérias pelo SIFT, PolyPhen 2, LRT e *Mutation Taster* em cada genoma (Figuras 33, 34, 35 e 36). Foram obtidas quatro redes de interação proteína-proteína com um nível de complexidade superior para os quatro genomas em estudo (Figuras 50, 52, 54 e 56).

H1

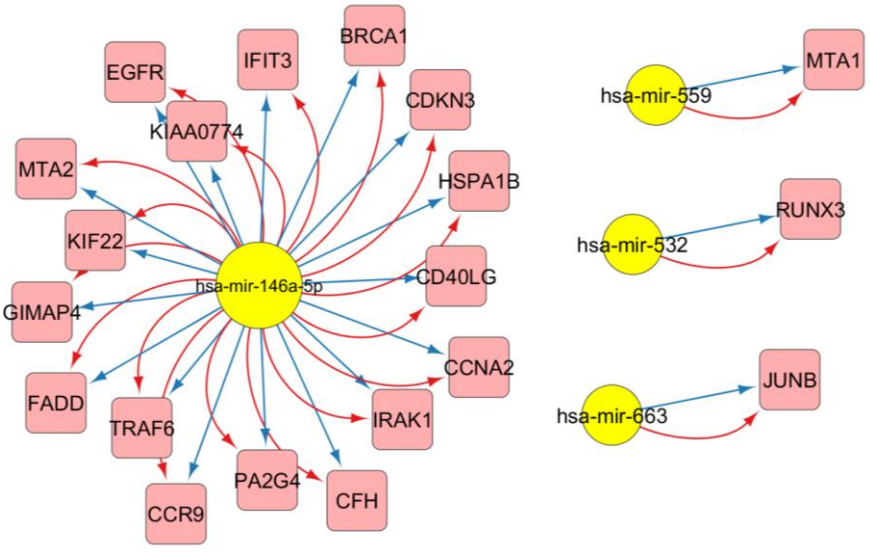


Figura 49: Representação efetuada através da plataforma Cytoscape com aplicação CyTargetLinker, dos miRNAs alterados no genoma H1, hsa-mir-146a-5p, hsa-mir-559, hsa-mir-532 e hsa-mir-663, com os respetivos genes alvo, resultantes da informação descrita na miRTarBase versão 4.5 (seta vermelha) e na Microcosm versão 5 (seta azul).

H1

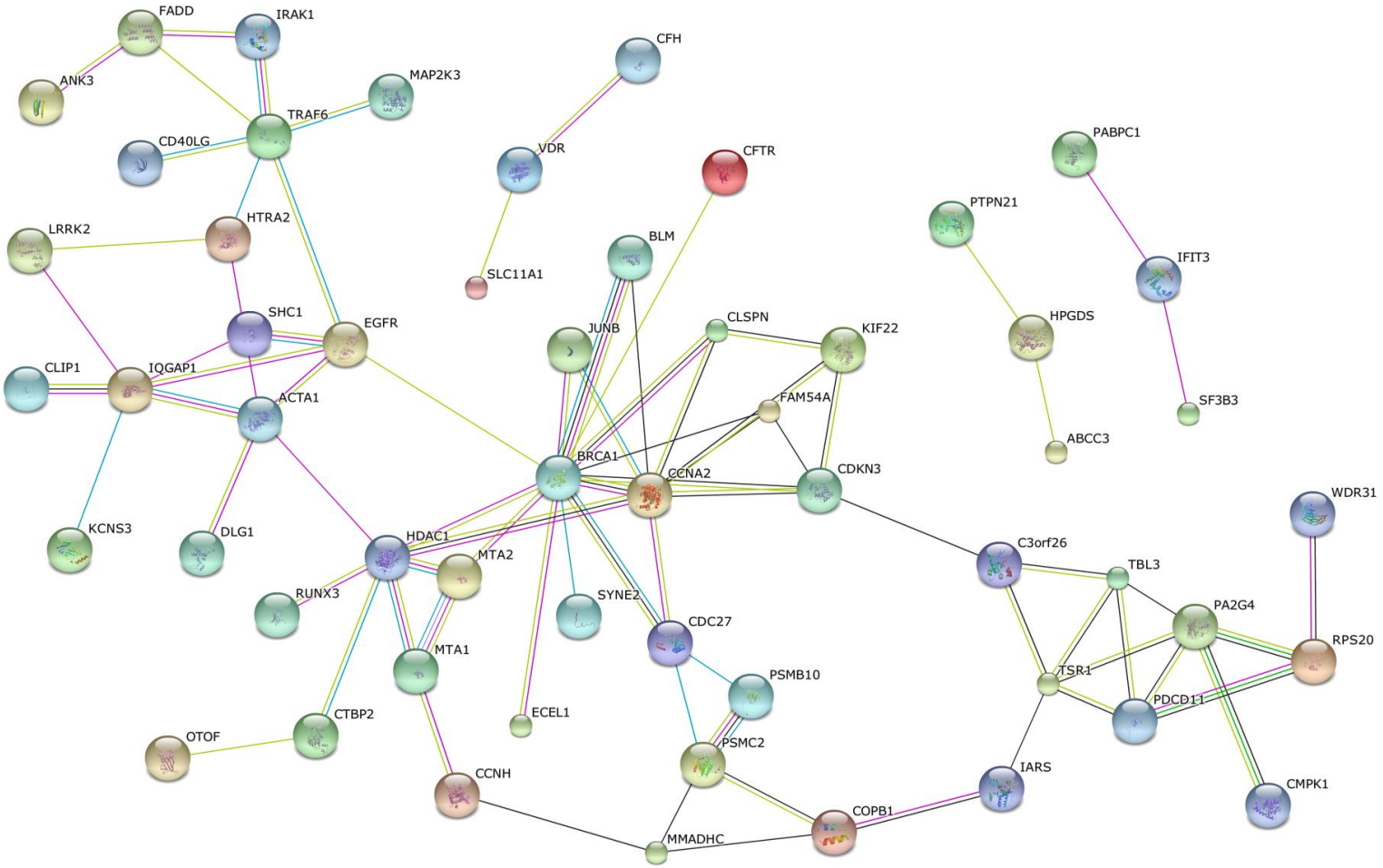


Figura 50: Análise das interações proteína-proteína efetuada pelo STRING, do produto dos genes codificantes com variantes identificadas no SIFT, PolyPhen 2, LRT e *Mutation Taster*, e o produto dos genes alvo de miRNAs alterados no genoma H1.

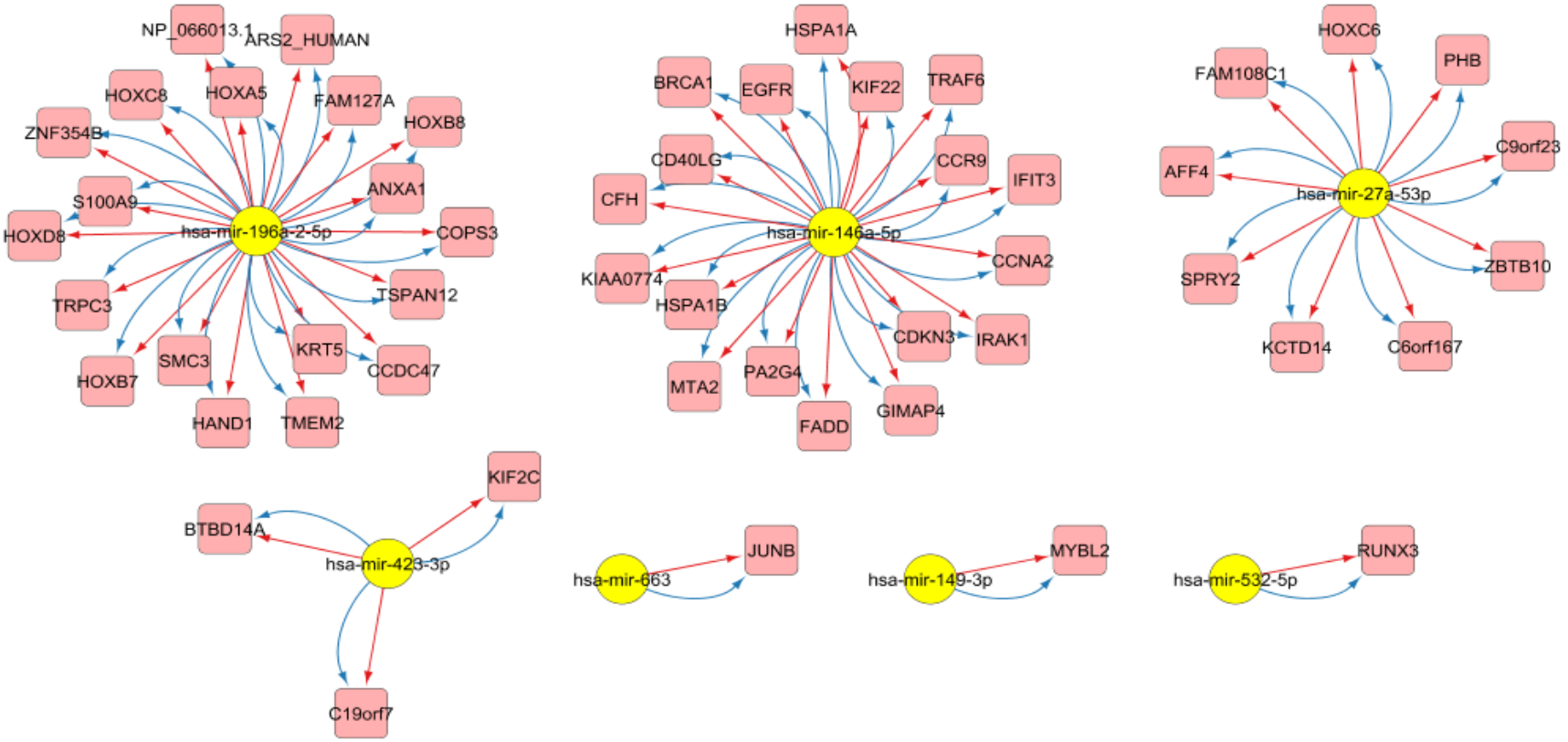


Figura 51: Representação efetuada através da plataforma Cytoscape com aplicação CyTargetLinker, dos miRNAs alterados no genoma H2, hsa-mir-196a-2-5p, hsa-mir-146a-5p, hsa-mir-27a-53p, hsa-mir-423-3p, hsa-mir-663, hsa-mir-149-3p e hsa-mir-532-5p, com os respetivos genes alvo, resultantes da informação descrita na miRTarBase versão 4.5 (seta vermelha) e na Microcosm versão 5 (seta azul).

H2

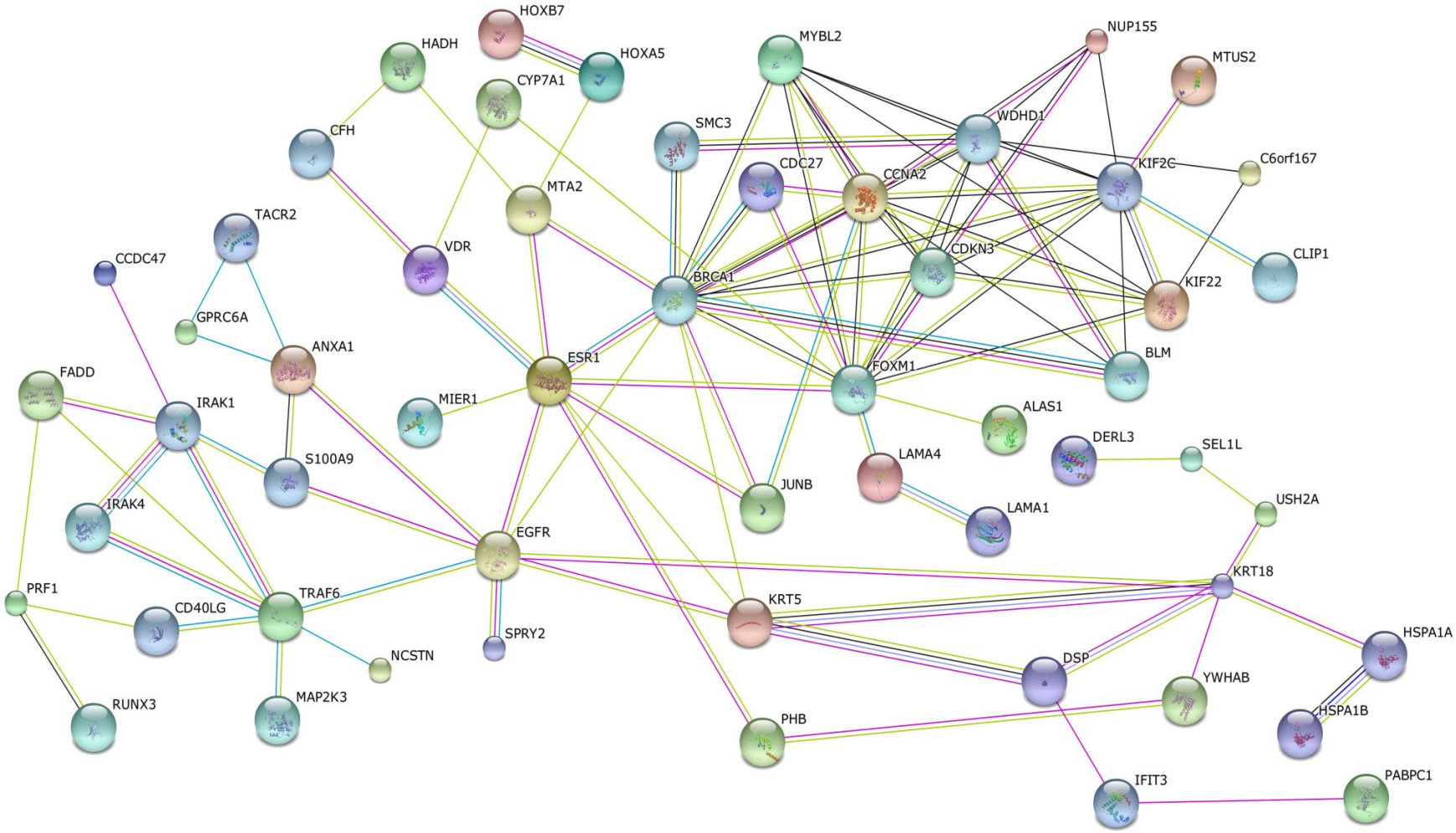


Figura 52: Análise das interações proteína-proteína efetuada pelo STRING, do produto dos genes codificantes com variantes identificadas no SIFT, PolyPhen 2, LRT e *Mutation Taster*, e o produto dos genes alvo de miRNAs alterados no genoma H2.

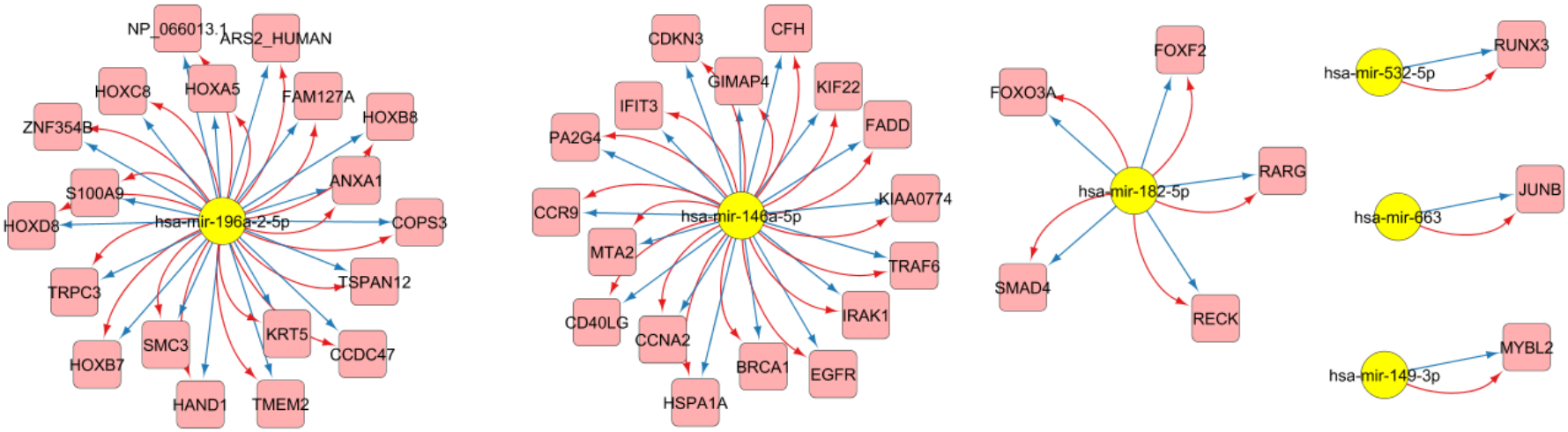


Figura 53: Representação efetuada através da plataforma Cytoscape com aplicação CyTargetLinker, dos miRNAs alterados no genoma H3, hsa-mir-196a-2-5p, hsa-mir-146a-5p, hsa-mir-182-5p, hsa-mir-532-5p, hsa-mir-663 e hsa-mir-149-3p, com os respetivos genes alvo, resultantes da informação descrita na miRTarBase versão 4.5 (seta vermelha) e na Microcosm versão 5 (seta azul).

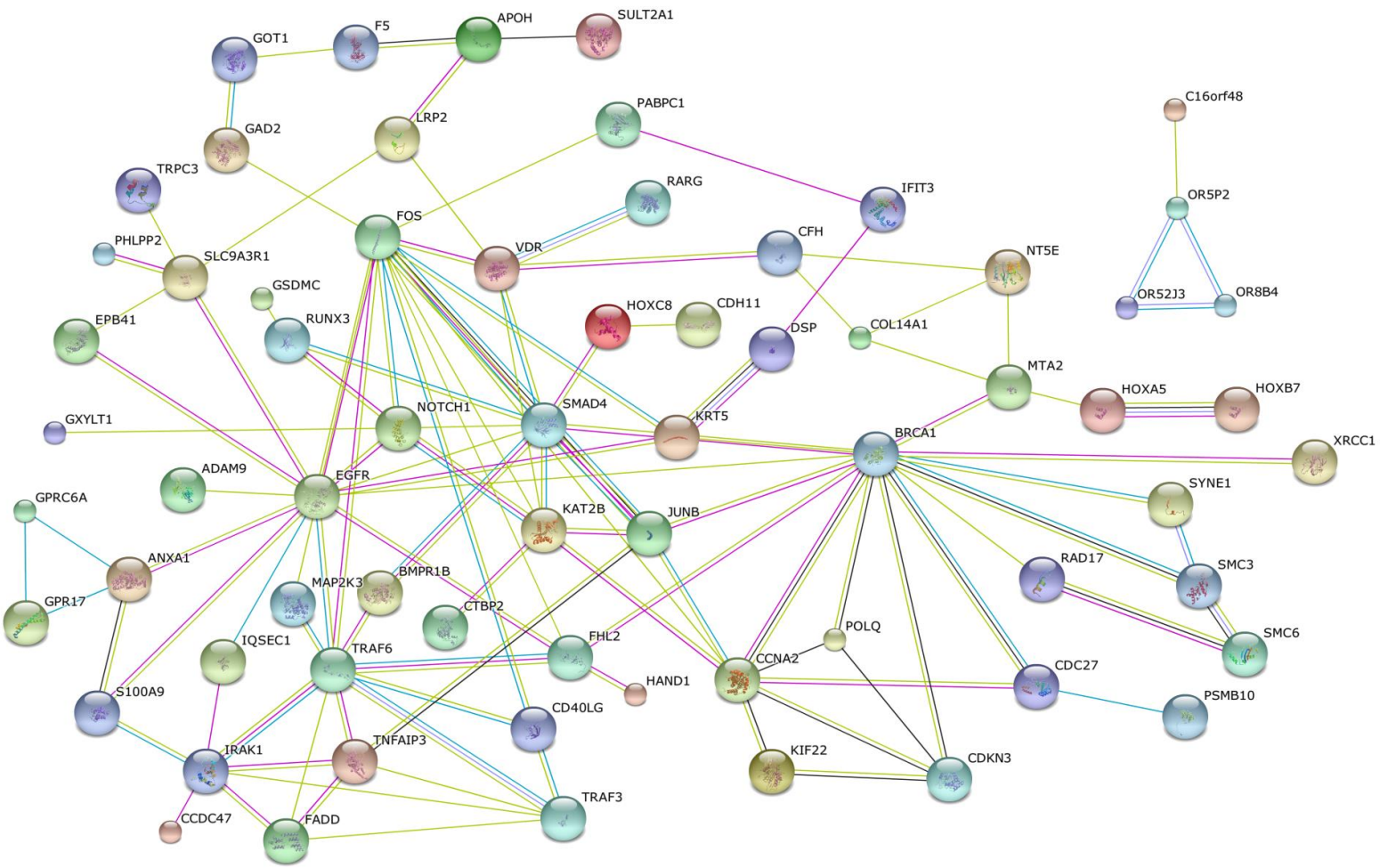


Figura 54: Análise das interações proteína-proteína efetuada pelo STRING, do produto dos genes codificantes com variantes identificadas no SIFT, PolyPhen 2, LRT e *Mutation Taster*, e o produto dos genes alvo de miRNAs alterados no genoma H3.

H4

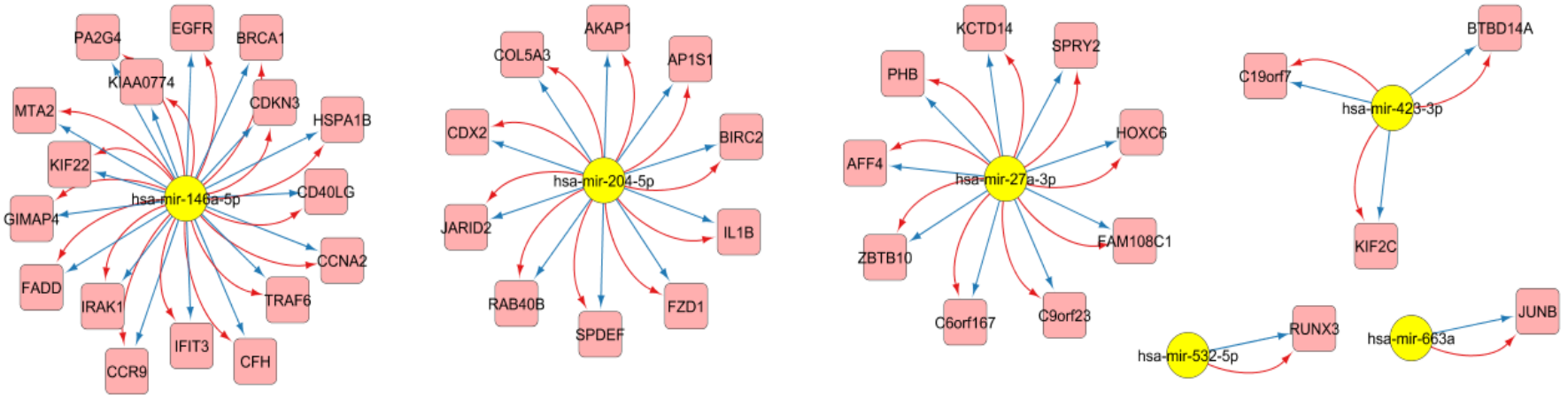


Figura 55: Representação efetuada através da plataforma Cytoscape com aplicação CyTargetLinker, dos miRNAs alterados no genoma H4, hsa-mir-146a-5p, hsa-mir-204-5p, hsa-mir-27a-3p, hsa-mir-423-3p, hsa-mir-532-5p e hsa-mir-663a, com os respetivos genes alvo, resultantes da informação descrita na miRTarBase versão 4.5 (seta vermelha) e na Microcosm versão 5 (seta azul).

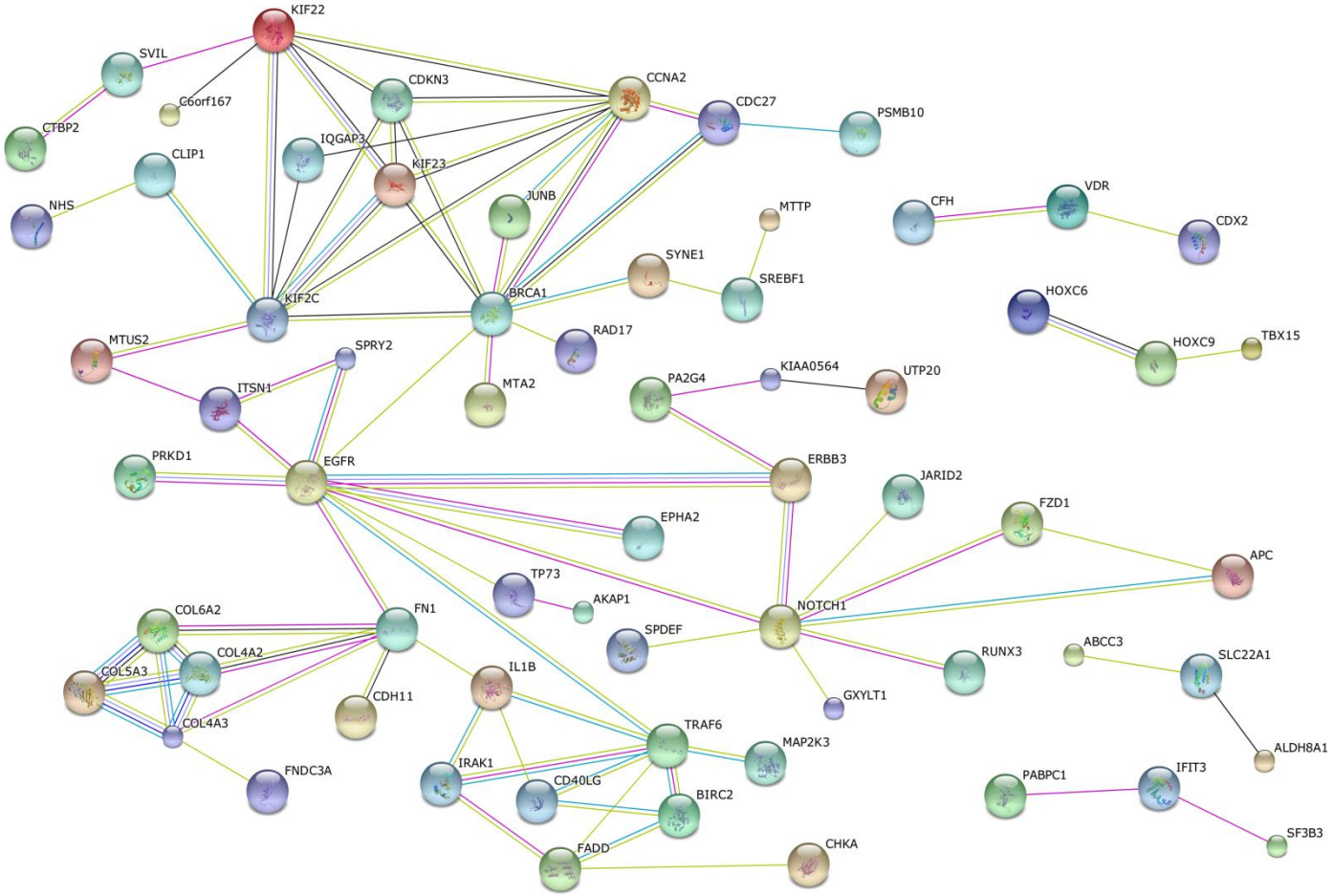


Figura 56: Análise das interações proteína-proteína efetuada pelo STRING, do produto dos genes codificantes com variantes identificadas no SIFT, PolyPhen 2, LRT e *Mutation Taster*, e o produto dos genes alvo de miRNAs alterados no genoma H4.

6. Ancestralidade

6.1. Estudo da ancestralidade genética dos quatro genomas

Foi realizado o estudo da ancestralidade genética para cada um dos quatro genomas em estudo, correspondentes a quatro indivíduos de etnia auto declarada portuguesa. Os resultados obtidos pela PCA estão representados na figura 57, correspondendo o eixo dos xx à PC1 e o eixo yy à PC2. A informação genética utilizada para caracterizar cada população representada na análise encontrava-se presente no HGDP e a caracterização de cada população a nível continental encontra-se na Tabela 13. É possível denotar na figura 57, a separação das populações continentais, sendo notória a presença dos genomas H1, H2, H3 e H4 na região ocupada pelos indivíduos das diferentes populações do continente europeu. Da representação resultante da análise da ancestralidade a nível continental foi possível evidenciar a distância da maioria das populações africanas das restantes populações mundiais tidas em consideração no estudo.

Foi efetuada uma análise às populações intracontinentais, relativamente às populações presentes no continente europeu, a qual evidenciou que a representação dos quatro genomas de indivíduos portugueses ocorreu nas proximidades com as populações da Rússia (Russian; Adygei), de Itália (Italian; Sardinian; Tuscan), da Escócia (Orcadian), de Espanha (Basque) e de França (French) (Figura 58). Algumas populações da Ásia ocidental aproximaram-se das populações europeias, como é o caso da população Druze.

Tabela 13: Posicionamento a nível continental das populações utilizadas no estudo da ancestralidade e presentes no HGDP.

Populações utilizadas no estudo da ancestralidade		
África		BantuKenya; BantuSouthAfrica; Biaka pygmy; Mandenka; Mbuti pygmy; Mozabite; San; Yoruba.
Ásia	**Ásia ocidental**	Bedouin; Druze; Palestinian.
	Ásia central e do Sul	Balochi; Brahui; Burusho; Hazara; Kalash; Makrani; Sindhi; Uyghur; Pathan.
	Ásia oriental	Cambodian; Dai; Daur; Han (North China); Hezhen; Japanese; Lahu; Miao; Mongolia; Naxi; Orogen; She; Tu; Tujia; Xibo; Yakut; Yi.
América		Karitiana; Maya; Pima; Surui.
Europa		Adygei; Basque; French; Italian; Orcadian; Russian; Sardinian; Tuscan.
Oceânia		Melanesian; Papuan.

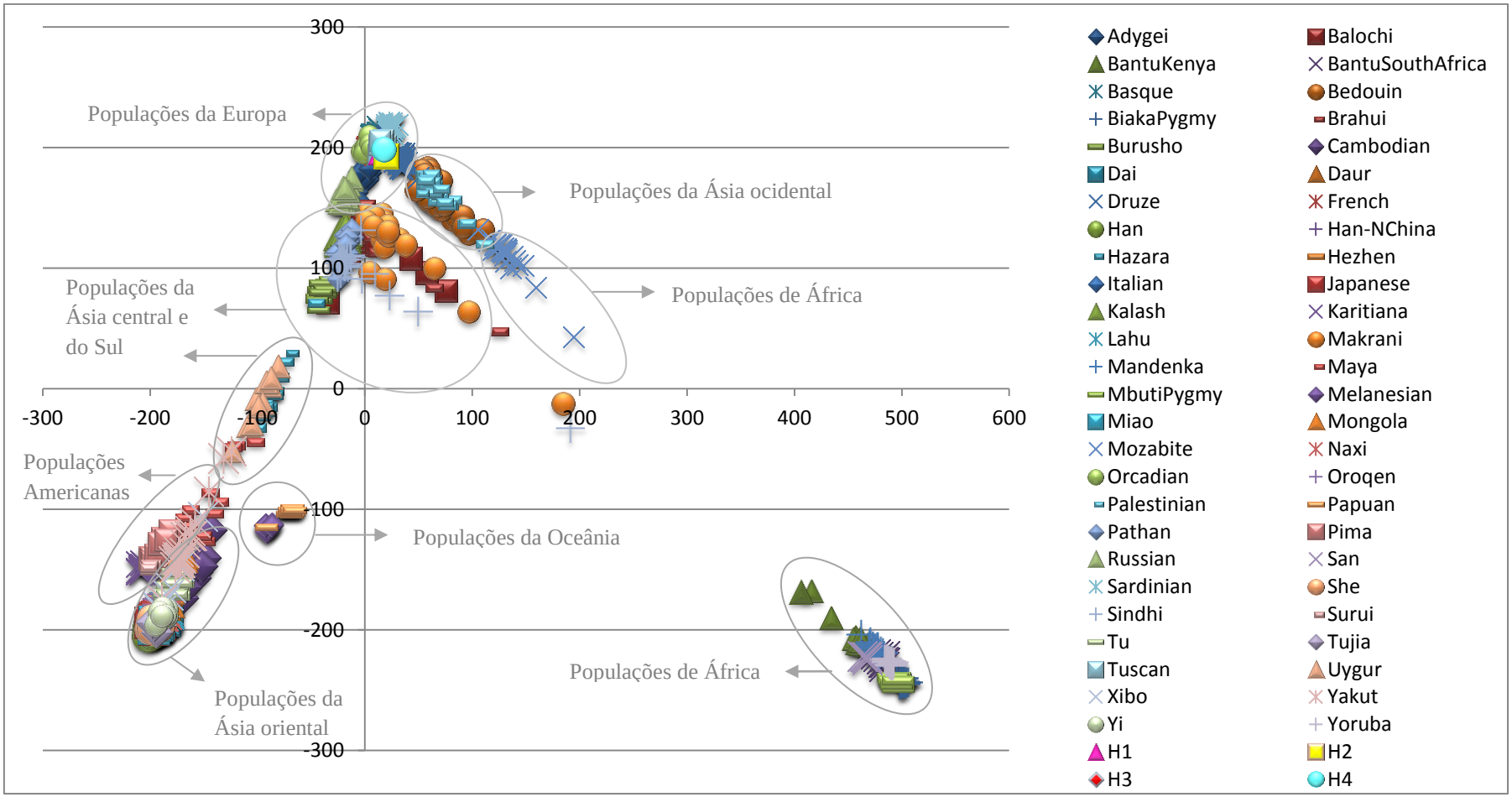


Figura 57: Representação da PCA efetuada para o estudo da ancestralidade genética global, realizada pela ferramenta LASER. Representação de todas as populações analisadas e caracterização continental das mesmas.

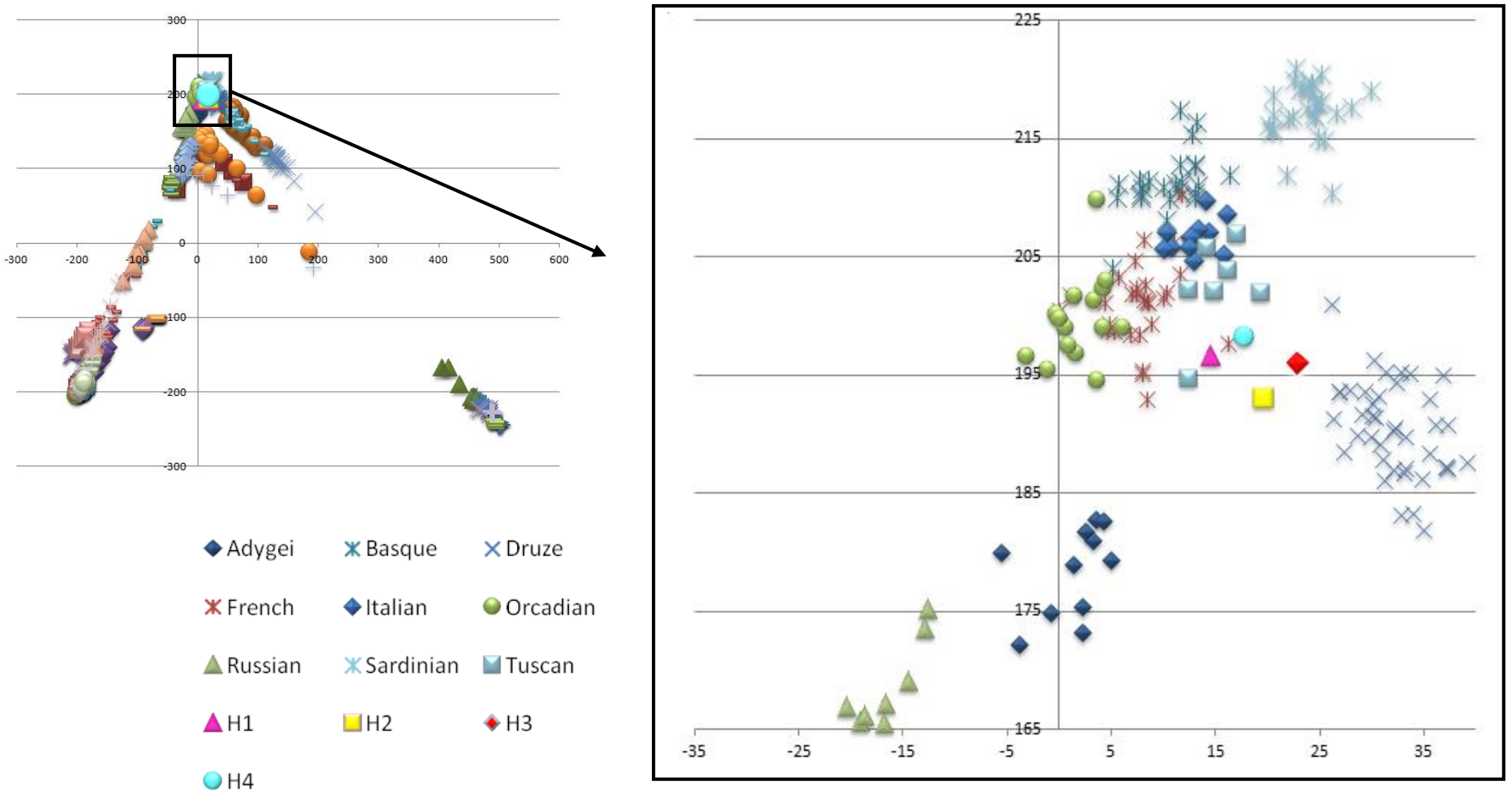


Figura 58: Análise detalhada da região ocupada pelas posições referentes às populações europeias em estudo, com particular destaque para a representação dos quatro genomas de indivíduos portugueses sequenciados (H1, H2, H3 e H4).

Discussão

A possibilidade, conferida pelas tecnologias NGS, de caracterizar o genoma humano com dados exatos e na sua totalidade, de forma a compreender mecanismos celulares e moleculares que ligam a informação presente no genoma a efeitos fenotípicos, norteou os objetivos do presente estudo onde sequenciámos os genomas completos de quatro indivíduos portugueses saudáveis (H1, H2, H3 e H4), pela tecnologia Illumina. Estes são, tanto quanto sabemos, os primeiros quatro genomas de indivíduos portugueses, sequenciados e analisados na sua totalidade.

Das *reads* geradas na sequenciação das bibliotecas genómicas, 97,17% correspondente ao genoma H1, 97,92% correspondente ao genoma H2, 98,08% correspondente ao genoma H3 e 97,79% correspondente ao genoma H4, foram mapeadas contra o genoma de referência hg19. Do mapeamento resultou, nos quatro genomas, uma cobertura elevada média de ≈34x. O valor médio aproximado de 2% de *reads* não mapeadas, em cada genoma, poder-se-á dever a fatores como: o método de sequenciação e o respetivo nível de qualidade associado ao produto deste; ao processo de mapeamento e às limitações de *software;* a diferenças genéticas entre etnias (genomas portugueses e o genoma de referência hg19); e às diferenças interindividuais, as quais podem ter uma variabilidade capaz de inflacionar a fração não mapeada. Os parâmetros de qualidade obtidos para cada genoma atestaram a confiança com que se prosseguiu a análise dos quatro genomas sequenciados. O *score* de qualidade médio por base em cada posição nas *reads* obtidas, para cada genoma, apresenta um nível de qualidade superior a Q30, denotando-se apenas nas últimas posições um decréscimo do *score* de qualidade, o que é esperado, uma vez que na maioria das plataformas, incluindo na plataforma Illumina, a identificação de bases sofre um decréscimo progressivo da qualidade ao longo da corrida de sequenciação. A qualidade média por *read* foi muito satisfatória, tendo os genomas H1, H3 e H4, apresentado um *Phred score* de ~Q30, decorrendo daqui uma exatidão de base de 99,9% (141). O genoma H2 apresentou um *Phred score* de ~40Q, de onde resulta uma exatidão de base de 99,99% (Tabela 1). Esta garantia dada pela análise da qualidade do produto da sequenciação de cada genoma, permitiu a análise das alterações de cada genoma relativamente à referência utilizada.

A comparação da sequência de nucleótidos de cada genoma e a sequência do genoma de referência permitiu identificar bases variantes para a análise da especificidade individual e populacional. No genoma H1, foram identificadas um total de 4.180.666 alterações, divididas por 3.826.898 SNPs e 353.768 INDELs; no genoma H2, o total das 4.935.137 variantes foram divididas por 4.398.887 SNPs e 536.250 INDELs; no genoma H3, das 4.854.704 variantes identificadas, 4.350.872 foram SNPs e 503.832 foram INDELs; e no genoma H4, o total de 4.376.224 alterações dividiu-se por 4.010.775 SNPs e 365.449 INDELs. Dos valores apresentados, torna-se claro que existem valores próximos de SNPs e INDELs entre os quatro genomas em estudo, evidenciando ainda que os nossos resultados acompanham a tendência dos valores de outros genomas individuais anteriormente descritos (42, 142), como é exemplo o genoma do primeiro indivíduo turco, cuja análise resultou em 4.251.434 variantes pontuais, divididas em 3.537.794 SNPs e 713.640 INDELs (43).

O estudo do padrão de distribuição destas alterações pelo genoma, bem como a análise da densidade de variantes em determinadas regiões genómicas, é um desafio atual pela informação sobre processos de seleção natural, evolução e de funcionalidade genómica que fornece. Analisando a forma como o total das variantes se distribuiu pelos cromossomas nucleares de cada genoma estudado, é possível notar a proporcionalidade que existe entre o número absoluto de variantes por cromossoma e o comprimento em número de pb que cada cromossoma apresenta. Esta proporcionalidade repercute-se nos quatro genomas (Tabelas 3 e 4). Como era esperado, como principal forma de variação do genoma, o número de SNPs nos quatro genomas foi notoriamente superior ao número de INDELs. Montgomery e seus colaboradores, na análise que efetuaram à distribuição e impacto de INDELs, identificados em 179 genomas humanos, observaram que embora 43% a 48% dos INDELs ocorram em cerca de 4,03% do genoma, nos chamados *hotspots* de INDELs, nos restantes 96% do genoma, a prevalência de INDELs é cerca de 16 vezes mais baixa que os SNPs (143). Apesar dos INDELs serem a segunda forma mais abundante de variação genética, o entendimento atual acerca da sua origem e dos seus efeitos funcionais carece de conhecimento. O principal desafio a esse estudo, prende-se com o facto de, quer as deleções quer as inserções, serem muito sensíveis a erros de alinhamento das *reads*, aquando da montagem do genoma (144), facto que merece a nossa precaução relativamente à caracterização de INDELs efetuada nos genomas em análise.

Já no que concerne à principal forma de variação genética pontual, a densidade de SNPs nos cromossomas nucleares foi avaliada através do rácio entre o número de bases variantes e o número de bases que constitui cada cromossoma do genoma de referência (Tabelas 3 e 4). Os cromossomas 19 e 21 apresentaram os rácios de variação mais elevados no conjunto dos resultados referentes aos quatro genomas, enquanto o rácio mais baixo foi identificado, unanimemente, no cromossoma X. Calculado o rácio médio de variação ao longo de cada genoma obteve-se o valor de 0,12%, 0,14%, 0,14% e 0,13%, no genoma H1, H2, H3 e H4, respetivamente. Este rácio revela, aliás, um valor expectável tendo em consideração que se estima que dois indivíduos aleatoriamente escolhidos tenham 99,9% de sequências idênticas (45). Não obstante, a variação na densidade de SNPs ao longo do genoma reflete pequenas diferenças na distribuição dos mesmos. Este facto deve-se em parte, à já referida existência de *hotspots* de recombinação em determinadas regiões do genoma (145). O projeto HapMap permitiu identificar ≈33.000 *hotspots* de recombinação (146). Outro importante fator recai na pressão exercida pela seleção natural, que conduz à redução ao longo das gerações da presença de alterações em regiões suscetíveis de produzir um fenótipo deletério para o organismo humano (por exemplo, em genes), bem como à pressão seletiva exercida no sentido de fixar as variantes que não desempenham um papel na alteração do *fitness* do sistema biológico humano (por exemplo, nas regiões intergénicas) (147).

Na prossecução deste entendimento, importa também atender à distribuição das alterações pontuais pelas diferentes regiões genómicas, ao longo de cada genoma. Nos quatro genomas em estudo, ≈65% dos SNPs identificados estavam na região intergénica. No interior da região génica, os SNPs distribuíram-se maioritariamente pelas regiões intrónica, exónica e pelas regiões flanqueadoras 3‘UTR e 5’UTR, numa percentagem de ≈ 31%, ≈ 0,6%, ≈ 0,6% e ≈ 0,1%, respetivamente, em cada um dos quatro genomas (Figuras 12, 14, 16 e 18). Relativamente à distribuição dos INDELs pelas diferentes regiões genómicas, a proporção nas distribuições manteve-se idêntica ao verificado nos SNPs de cada genoma (Figuras 13, 15, 17 e 19). Não obstante da relação de grandeza que cada uma das regiões enunciadas tem no genoma humano, vários estudos evidenciaram já a importância da pressão que a seleção natural exerce na diferenciação da densidade de alterações ao longo das regiões, denotando-se a seleção positiva em regiões intergénicas, ou que se considera de menor importância funcional, e uma seleção purificadora em

regiões codificantes, de importância fenotípica elevada (148-150). Zhao e seus colaboradores sugerem a existência de seleção natural, nomeadamente de seleção purificadora, na modulação da densidade de SNPs ao longo do genoma humano (148). Os nossos valores são corroborados por iguais tendências de distribuição no primeiro genoma irlandês sequenciado e analisado com elevada cobertura, que apresenta 63% das suas variantes na região intergénica e os restantes 37% na região intragénica, os quais se dividem em 35% na região intrónica, 0,6% na região exónica, 0,6% na região 3'UTR e 0,1% na região 5'UTR (151).

Detendo-nos sobre o comprimento dos INDELs, nos quatro genomas em estudo, verificaram-se em maior número inserções com comprimento de + 1 base e deleções com comprimento de – 1 base (Figuras 23, 24, 25 e 26). Torna-se previsível que a alteração de um menor número de bases seja suscetível de causar menos prejuízo fenotípico e desequilíbrio estrutural da própria molécula de DNA, do que os INDELs de maior número de bases. A média do número de bases inseridas é de ≈2 pb e a média do número de bases deletadas é de ≈3 pb, nos quatro genomas portugueses, o que também foi verificado no genoma sequenciado do indivíduo asiático YH (37), ou no genoma do indivíduo indiano IGIB1 (142). Relativamente ao número de inserções e deleções, nos quatro genomas em estudo, o número de deleções, ≈ 53%, sobrepôs-se ao número de inserções, ≈ 47%. Embora alguns autores sugiram que as pequenas deleções são mais deletérias que as pequenas inserções (152, 153), outros têm argumentado que estas inserções são mais deletérias por aumentar a sequência passível de ser alterada, ao passo que as deleções removem sequências que poderão conter outras alterações deletérias. (154). De acordo com a teoria de seleção purificadora descrita, pelos valores obtidos parece-nos a segunda hipótese mais viável, embora carecendo de estudos mais direcionados para o determinar. Da análise efetuada, quer da densidade das alterações pelos cromossomas nucleares, quer da ocorrência destas nas diferentes regiões genómicas, é possível inferir que as variantes embora sejam consideradas aleatórias, uma vez que são consideradas fenómenos estocásticos, não são equiprováveis.

Atendendo ao tipo de SNPs distribuídos pelos cromossomas nucleares nos quatro genomas, é possível verificar em todos eles, um aumento da ocorrência de substituições de pirimidinas por pirimidinas, ou purinas por purinas, sobre a ocorrência de substituições de

pirimidinas por purinas, ou o contrário (Figura 20). Devido ao facto de existirem quatro possibilidades de Ts e oito possibilidades de Tv, o rácio esperado de Ts/Tv seria de 0,5 (155). No entanto, no genoma H1 o rácio Ts/Tv foi de 1,92; no genoma H2 de 1,87; no genoma H3 de 1,88; e no genoma H4 de 1,91. Assim, nos quatro genomas, o valor traduziu a superioridade das Ts em relação às Tv em todos os cromossomas, e em alguns (cromossoma 1 e 2) quase para o dobro do valor de Tv (Figura 20). O facto das substituições do tipo Ts envolverem a troca de nucleótidos estruturalmente mais semelhantes, isto é, a troca entre nucleótidos com bases orgânicas nitrogenadas de um único anel (citosinas e timinas), ou a troca de nucleótidos com bases orgânicas nitrogenadas de anel duplo (adenina e guanina), relativamente às substituições do tipo Tv, que envolvem a troca entre nucleótidos mais distintos, com a troca entre um nucleótido com um único anel e um nucleótido de anel duplo, poderá ser uma condicionante assinalável para explicar os valores obtidos. No entanto, o *bias* causado pelo aumento da ocorrência de Ts relativamente à ocorrência de Tv poderá ter como razão determinante o mecanismo molecular através do qual estas são geradas. Embora não haja uma resposta absoluta para a questão que envolve as causas das variantes genéticas, muito pela associação das variantes pontuais a processos quânticos e, portanto, com um princípio de incerteza irredutível, as Ts apresentam um fenómeno molecular bem definido associado à sua ocorrência, o processo de tautomerização (156). Cada uma das bases nucleotídicas existe em dois estados tautoméricos alternativos, sendo este um caso de isomeria funcional, em que os dois isómeros estão em equilíbrio dinâmico. Este equilíbrio é fortemente deslocado para o lado das estruturas convencionais, que representam os estados predominantes, e que permitem o correto emparelhamento das bases. De forma espontânea, e reversível, ocorrem desvios de tautomerização em que existe uma mudança da forma amino (mais estável) para a forma imino das bases nucleotídicas, ou a alteração espontânea da forma ceto (mais estável) para a forma enol dos nucleótidos, o que leva à alteração de estados das bases envolvidas conducentes ao emparelhamento destas com bases não canónicas, favorecendo a ocorrência de Ts (157). Dogan e seus colaboradores, na análise do primeiro genoma completo de um indivíduo turco, inferiram resultados semelhantes aos nossos, com um rácio Ts/Tv de 2,06 (43).

A distribuição das alterações ao longo de cada genoma humano, bem como o tipo de variante em questão, reflete o grau de evolução biológica e o fluxo de variações inter e

intrapopulacionais. Neste sentido, o número de variantes identificadas pela primeira vez assume um papel fundamental em qualquer estudo genómico, quer individual quer populacional. Nos quatro genomas estudados, em média ≈ 11% dos SNPs e ≈ 29% dos INDELs foram identificados pela primeira vez, isto é, não há registo destas variantes na base de dados dbSNP137 (Figura 27). Primeiramente destaca-se, em todos os genomas, um valor aproximadamente três vezes superior de INDELs novos, inflação esta que poderá estar relacionada com os erros no alinhamento, e consequente identificação incorreta de INDELs, anteriormente referidos; ou, numa segunda análise, com o menor volume de trabalhos científicos de identificação de INDELs, comparativamente aos estudos que têm por alvo a análise de SNPs. No entanto, quer os SNPs quer os INDELs identificados como novos, embora detenham um maior risco no que concerne à sua veracidade, são as variações que são passíveis de ser específicas do indivíduo e/ou da população em que se inserem, neste caso, da população portuguesa.

Shen e seus colaboradores caracterizaram as variantes genéticas identificadas após a sequenciação de 44 genomas humanos completos de indivíduos caucasianos com elevada cobertura, e reforçaram o entendimento de que a grande maioria das variantes identificadas pela primeira vez têm uma frequência alélica baixa, ou rara, as quais estão intimamente relacionadas com a suscetibilidade para a doença (158). Esta compreensão traduz-se na necessidade de estudar um maior número de genomas individuais com elevada cobertura, uma vez que grande parte das variantes com importância fenotípica podem assim estar por descobrir.

Da análise que efetuamos ao número de SNPs novos e conhecidos distribuídos pelos diferentes cromossomas dos quatro genomas verificámos, como era expectável, que o número de alterações conhecidas superou o número de variantes novas em quase todos os cromossomas (Figura 28). Exceção feita para o cromossoma Y, no qual se denota, em todos os genomas, um aumento dos SNPs novos relativamente aos SNPs já reportados ao dbSNP137. O cromossoma Y é frequentemente usado como marcador para estudar a história da demografia humana, devido às suas características distintas, nomeadamente à sua natureza haplóide e à ausência de recombinação, que permitem uma herança monoparental e que é essencial à genética populacional (159). O número elevado de alterações novas reflete a singularidade deste cromossoma, bem como o seu potencial na

caracterização individual e, muito provavelmente, populacional. Um estudo recente refuta a ideia assumida de que o cromossoma Y não é afetado por seleção natural, tal como se verifica noutros cromossomas (160). Sayres e colaboradores admitem que um fenómeno de seleção purificadora vai sendo exercido nas variantes do cromossoma Y, o que está concordante com a diminuição do rácio de variação observada nos nossos resultados (160).

Comparando os resultados obtidos com os descritos nos genomas individuais publicados nos últimos anos, é possível verificar que relativamente ao número de SNPs, conhecidos e novos, os nossos resultados acompanham as proporções, enquanto que para os INDELs, o número de INDELs novos tem vindo progressivamente a reduzir. O primeiro genoma de um individuo coreano apresentou 66,9% de INDELs novos (38); o primeiro genoma completo sequenciado de um individuo irlandês, apresentou 50,7% de INDELs novos (151); e o primeiro genoma completo de um individuo turco apresentou 22,0% de INDELs novos (43). Tal redução no número de INDELs novos deve-se, muito provavelmente, a melhorias na metodologia na identificação dos mesmos, bem como a um aumento dos estudos destas variantes, embora os valores ainda se mantenham altos comparativamente aos SNPs.

No que concerne à zigotia das variações, novas e conhecidas, identificadas em genomas anteriormente publicados, o número de alterações heterozigóticas é superior ao número de alterações homozigóticas, apresentando todos uma proporção de homozigóticos/heterozigóticos de ≈ 40%/60% (43, 142, 151). A mesma proporção foi verificada nas variações identificadas nos quatro indivíduos em estudo, tendo o genoma H1 e H4 apresentado uma proporção de homozigóticos/heterozigóticos de ≈ 40%/60% e os genomas H2 e H3 de ≈ 30%/70% (Figura 30).

A translação das características encontradas na análise das alterações genéticas identificadas nos quatro genomas, para a compreensão das mesmas ao nível do impacto fenotípico e da sua aplicabilidade na melhoria do sistema biológico humano individual, orienta estudos genómicos de larga escala como o nosso. No fundo, o objetivo é compreender de que forma a combinação das variantes genéticas de cada um dos seus indivíduos afetam a sua suscetibilidade pessoal a doenças e modulam a resposta terapêutica. Este princípio é aliás, o que sustenta a pretendida meta da medicina personalizada. Estando esta disciplina a emergir em Portugal, o nosso estudo mostra-se

pioneiro na obtenção da informação à escala genómica que para ela contribui. Neste sentido, após a caracterização dos elementos genéticos identificados, a análise de cada genoma centra-se na região codificante de cada um, no sentido de prever o efeito a jusante do produto deste. Na região exónica do genoma H1, foram identificados 11.662 SNPs não sinónimos, isto é, que provocam alteração de aminoácido; no genoma H2 foram identificadas 12.426 SNPs (Tabela 5); no genoma H3 12.459 SNPs; e no genoma H4 12.010 SNPs (Tabela 6). Foi sobre este conjunto de alterações, passíveis de causarem o maior impacto deletério, que recaiu a análise combinada das quatro ferramentas bioinformáticas, SIFT, PolyPhen2, LRT e *Mutation Taster*. Cada uma destas ferramentas atribuiu um *score*, classificando as variantes como deletérias ou neutras, mediante os seus parâmetros de previsão (Tabela 9). Liu e colaboradores referem que a combinação de resultados que derivam das diferentes ferramentas supracitadas conduz a conclusões com uma confiança reforçada, uma vez que estes utilizam algoritmos diferentes e recurso a informação diversificada (161). Assim, selecionadas as variantes previstas com potencial deletério pelos quatro *softwares* em simultâneo, foram identificados os genes que compreendiam as referidas variantes em cada genoma. Foram identificados 129 genes no genoma H1 e 116 genes em cada um dos genomas H2, H3 e H4 (Figuras 32). Este conjunto, acreditamos nós, terá uma maior probabilidade de representar os genes com maior suscetibilidade para a alterar o *fitness* do organismo de cada individuo estudado. A nossa seleção das variantes e dos respetivos genes é corroborada pela análise das MAF das variações em questão, tendo em média 93,6% e 97,4% destas uma MAF<5%, no projeto dos 1000 Genomas e no ESP, respetivamente. Como referido anteriormente, a conceção de que as MAF mais baixas (MAF<5%) estão associadas a uma contribuição major na doença, atesta a seleção (52). Assim como o facto de nos quatro genomas ≈ 90% das variações selecionadas em cada um estarem presentes em regiões conservadas do genoma, onde se espera que as alterações presentes sejam mais deletérias (162).

Da observação dos genes que compreendem as variantes selecionadas decorre ainda o facto da classe relativa à função biológica dos genes mais alterados ser aquela cujos produtos proteicos estão envolvidos no processo metabólico. O metabolismo celular refere-se ao conjunto de todas as reações químicas que ocorrem nas células e, por isso, apresenta um grande conjunto de proteínas expressas a partir de vários genes presentes no exoma total. No entanto, são também estas as reações responsáveis pelos processos de síntese e

degradação dos nutrientes na célula e que constituem a base da vida, permitindo o crescimento e a reprodução das células, exacerbando assim o impacto da sua alteração no organismo humano (163).

A suscetibilidade baseada em características genéticas para a alteração do sistema biológico de cada individuo recai não só na análise da predisposição para a ocorrência de doença, como também, na variabilidade associada à resposta terapêutica, ou seja, alterações em genes que codificam enzimas metabolizadoras de medicamentos, recetores e proteínas que afetam a absorção e transporte de princípios ativos. Esta variabilidade explica que diferentes pacientes, com a mesma condição, expostos à mesma dose de determinada medicação, denotem uma eficácia e uma suscetibilidade a efeitos adversos marcadamente diferentes.

Detendo-nos sobre o genoma H1, identificou-se uma alteração no gene *COL4A3* (rs190598500), associada à síndrome de *Alport* (164). Esta síndrome, com um padrão de hereditariedade autossómico recessivo, é caracterizada por uma alteração na síntese de colagénio, nomeadamente colagénio tipo IV, o que provoca uma perda progressiva da função renal e auditiva. A proteína do gene que continha esta variante relaciona-se com a proteína do gene *HABP2* (Anexo 7), estando esta envolvida na adesão celular e ligação a proteínas da matriz extracelular, nomeadamente ao colagénio tipo IV (165). No gene *HABP2* foi ainda identificada uma variante (rs7080536) descrita como sendo um fator de risco independente para a estenose da carótida (166) e para o tromboembolismo venoso (167). Neste genoma foi identificada uma variante (rs16891982) no gene *SLC45A2*, que codifica um transportador proteico que medeia a síntese de melanina. Esta proteína tem como função regular a pigmentação da pele, olhos e cabelos, e a alteração identificada tem sido descrita como fator determinante da cor da pele clara dos europeus (168). Esta variante é por isso considerada como marcador de ancestralidade (169). Não é de estranhar, portanto, que a mesma alteração identificada no genoma H1 tenha sido identificada também no genoma H2 e H3. Esta variante tem sido, nos últimos anos, definida como fator genético de proteção da ocorrência de melanoma maligno, nomeadamente em estudos realizados na população espanhola e francesa (170, 171). Contudo, um outro estudo envolvendo populações europeias, associa esta alteração à ocorrência de carcinoma basocelular e do carcinoma das células escamosas (172). O

genoma H1 apresentou variantes nos genes *SHC1*, *HDAC1* e *CTBP2*, e embora não estejam descritas associações destas a fenótipo, as proteínas dos genes referidos estão descritas como associadas à via da leucemia mieloide crónica (via KEGG:hsa05220), encontrando-se no mesmo *cluster* de interação previsto pelo STRING (Figura 33). Uma outra alteração (rs35687416), identificada no gene *CMPK1* deste genoma, está associada à alteração da resposta terapêutica à gemcitabina, um fármaco usado como quimioterapia para o tratamento de tumores sólidos, que atua durante a replicação de DNA induzindo a apoptose (173). No genoma H1, foi ainda identificada uma variante (rs33958906), no gene *LRRK2*, que é associada à ocorrência da doença de Parkinson (174). A proteína expressa pelo gene *LRRK2* interage com a proteína do gene *HTRA2* (Figura 33), ambas descritas como estando associadas à doença de Parkinson (via KEGG:hsa05012), embora a alteração identificada no gene *HTRA2* não esteja descrita como associada a fenótipo (175). O gene *HTRA2* está por sua vez relacionado com a proteína do gene *TRAF6* no complexo recetor CD40 (Figura 50), a expressão do qual é regulada pelo miRNA has-mir-146a-5p que se mostrou alterado neste indivíduo (Figura 49). O genoma H1 é ainda portador de uma variante (rs35077384) no gene *ZFYVE27*, cuja proteína está envolvida no transporte vesicular, e que quando alterada conduz a um tráfego neuronal intracelular afetado no trato corticoespinal. Esta alteração é consistente com a ocorrência de paraplegia espástica do tipo hereditário, uma doença de hereditariedade autossómica dominante, que pode resultar na disfunção dos nervos dos membros inferiores com uma rigidez e contração (espasticidade) progressiva (176).

No genoma H2, além da alteração (rs16891982) identificada no gene *SLC45A2*, e cujas consequências foram anteriormente mencionadas, foi identificada uma variante (rs1799807) no gene *BCHE*, que conduz à diminuição da expressão da butirilcolinesterase (177). A deficiência de butirilcolinesterase é uma doença metabólica que conduz a apneia prolongada após a utilização de medicamentos anestésicos, sendo esta uma deficiência de padrão autossómico recessivo e que apresenta uma maior prevalência na população caucasiana. Dado o seu caracter assintomático até à exposição a um medicamento anestésico, a identificação de portadores destas variantes tem uma enorme importância no campo da prevenção das suas manifestações. Neste genoma, foi também identificada uma alteração no gene *DSP* que não está descrita na bibliografia como associada a fenótipo. No entanto a proteína do gene *DSP* apresentou interação com a proteína do gene *IFIT3* pela

análise STRING (Figura 52). O gene *IFIT3* embora não tenha sido identificado com uma variante não sinónima na nossa análise é regulado pelo miRNA hsa-mir-146a-5p, o qual se demonstrou estar alterado neste genoma (Figuras 51). A análise STRING mostrou também a interação entre a proteína do gene *DSP* e a proteína do gene *KRT18*, gene este com uma alteração identificada embora não estando associada a fenótipo, e com a proteína do gene *KRT5* (Figura 52). O gene *KRT5* embora não tenha apresentado alterações no genoma H2, é regulado pelo miRNA alterado hsa-mir-196a-2-5p (Figuras 51). A variante (rs35947132) identificada no gene *PRF1* neste genoma está associada à linfohistiocitose hemofagocítica familiar (178). O gene *PRF1* codifica a porfirina 1, a qual mostrou interação com as proteínas do gene *RUNX3*, do gene *FADD* e do gene *CD40LG* (Figura 52), estando a expressão de todas estas proteínas regulada pelos miRNAs identificados com alterações, hsa-mir-532-5p e hsa-mir-146a-5p (Figura 51).

O genoma H3, concomitantemente com o verificado no genoma H1 e H2, apresentou a alteração (rs16891982) no gene *SCL45A2* já mencionada. Uma outra alteração (rs143813189) foi identificada no gene *TRAF3* deste genoma, a qual está descrita como sendo um fator de risco para a ocorrência de encefalite em indivíduos com o vírus Herpes simplex 1, suscetibilidade que é, no entanto, mais preponderante em crianças (179). Foi também identificado como alterado o hsa-mir-146a-5p, descrito como regulador do gene *TRAF6* (Figura 53). Os genes *FADD*, *EGFR* e *SMAD4*, nos quais não se observaram alterações, são regulados por miRNAs identificados como alterados neste genoma, os dois primeiros pelo hsa-mir-146a-5p e o último pelo has-mir-182-5p (Figuras 53). Os genes *TRAF6*, *FADD*, *EGFR*, *SMAD4* regulados por miRNAs identificados como alterados, e os genes *FOS* e *CTBP2*, alterados com variantes não associadas a fenótipo, apresentaram interação das suas proteínas pela análise STRING (Figura 54), estando descritas como associadas a vias tumorais (via KEGG: hsa05200). Neste genoma, observou-se ainda a alteração (rs1799782) no gene *XRCC1*, que modifica a resposta terapêutica à cisplatina e à ciclofosfamida, ambos fármacos citotóxicos, usados no tratamento de linfomas, leucemias e tumores sólidos (180).

No genoma H4, a variante (rs190598500) presente no gene *COL4A3*, foi identificada, tal como já havia sido no genoma H1, anteriormente discutido. Pelo facto de esta variação ter uma MAF<1% na população mundial e ter sido identificada em dois dos

quatro genomas estudados mereceu especial atenção. Este resultado pode dever-se a: erro de sequenciação, o que neste caso não nos parece o mais provável, dada a validação manual da variante no ficheiro BAM através do *software* IGV; ao facto desta variante poder ser comum na população portuguesa; ou devido à possibilidade de se encontrar mal referenciada nas bases de dados. No genoma H4 foi também identificada a alteração (rs34116584) no gene *AGXT* que está associada à ocorrência de hiperoxalúria primária do tipo 1. Esta doença é causada por uma deficiência da enzima hepática peroxissomal, a L-alanina-glioxilato aminotransferase, a qual conduz à depuração de oxalato do organismo, que ao ser acumulado leva a urolitíase e a nefrocalcite (181). Esta variante está também associada à alteração da metabolização dos fármacos flourouracil e oxoliplatina, quimioterápicos usados no tratamento de diversos tumores, bem como de leucovorina, um fármaco usado em combinação com quimioterápicos no sentido de aumentar a eficácia do tratamento e ser quimioprotetor (182). Uma outra variante (rs34059508) identificada no gene *SLC22A1*, está associada à modificação da resposta terapêutica à metformina, um antidiabético oral, ao tropiretron, um antiemético prescrito para tratar efeitos colaterais de quimioterapias, e ao ondansetron, um antiemético e antivertiginoso que atua ao nível do sistema nervoso central (183). Embora não sejam conhecidas associações diretas entre as alterações identificadas em vários genes e fenótipo, a suas proteínas estão associadas a vias tumorais, nomeadamente as proteínas dos genes *NOTCH1*, *COL4A3* e *COL4A2*, que têm funções de regulação negativa da angiogénese (GO: 0016525), e as proteínas dos genes *COL4A2* e *FN1* associadas à ocorrência de cancro do pulmão (KEGG: hsa05222) (Figura 36). Estas proteínas apresentaram interação com as proteínas dos genes *EGFR*, *TRAF6*, *BIRC2*, *FADD* e *FZD1* pela análise STRING (Figura 56), também elas envolvidas em vias tumorais (KEGG: map05200). Embora estes genes não tenham apresentado variantes não sinónimas neste genoma, são regulados pelos miRNAs hsa-mir-146a-5p e has-mir-204-5p que foram identificados como estando alterados no genoma H4 (Figura 55).

A análise de suscetibilidades genéticas para determinadas condições, efetuada em relação à região codificante de cada genoma, salienta a majoração do impacto das variações não sinónimas que existe pela adição de dados de farmacogenómica e de elementos não codificantes de elevado poder de regulação da expressão génica, como a análise de miRNAs. Assim, esta análise global eleva a capacidade de compreender o mecanismo de ação de mais elementos intervenientes no *fitness* do organismo.

A abordagem utilizada no presente estudo permitiu caracterizar predisposições genéticas em cada um dos indivíduos cujo genoma foi sequenciado e colocou em evidência as potencialidades da medicina personalizada. Todavia, a atual inacessibilidade da genómica na rotina clínica leva a que o premente desafio passe pelo estudo das características genómicas populacionais, as quais permitem estudar suscetibilidades genéticas passíveis de serem específicas de uma população e auxiliam na construção da própria história ancestral desta. Na prossecução deste objetivo, foram analisados os SNPs comuns aos quatro genomas em estudo, tendo sido identificados 1.697.587 SNPs. Análises semelhantes efetuadas em genomas anteriormente sequenciados mostram valores próximos aos identificados no nosso estudo, como foi visível na comparação de SNPs entre o genoma de Watson, o genoma de Venter e o primeiro genoma de um indivíduo coreano, SJK, com um total de 1.254.570 SNPs comuns (38). A distribuição pelos cromossomas nucleares dos SNPs comuns aos quatro genomas portugueses evidenciou a mesma proporcionalidade relativamente à distribuição do total de SNPs de cada genoma individual (Tabela 10). Este facto relaciona-se, novamente, com o comprimento de pb de cada cromossoma. No que concerne à densidade de SNPs comuns, resultante da razão entre o número de SNPs e o número de pb de cada cromossoma, o cromossoma 21 apresentou o maior rácio, com 0,08%, a par do que já havia sido verificado em cada genoma individual (Tabela 10). A distribuição dos SNPs comuns pelas diferentes regiões genómicas foi concordante com a distribuição verificada em cada genoma individual (Figura 40). O tipo de SNPs comuns observados respeitou a tendência verificada em cada genoma individual, onde o número de Ts foi superior ao número de Tv partilhados, em cada cromossoma, tendo o rácio Ts/Tv o valor de 1,87 (Figura 41). Este resultado denota o enviesamento da ocorrência de Ts, tal como havia sido verificado em cada genoma e, em concordância com os resultados obtidos em toda a caracterização dos SNPs partilhados.

Atendendo ao valor de 7,4% dos SNPs novos que foram comuns aos quatros genomas, este mostrou-se inferior à fração de SNPs novos verificada nos genomas individualmente, de ≈11% em cada (Figuras 27 e 44). Assim, acreditamos que os 7,4% de SNPs novos comuns aos quatro apresentam uma maior probabilidade de serem específicos da população portuguesa, sendo os restantes SNPs novos identificados individualmente, os que apresentam maior probabilidade de serem específicos do individuo. Considerando a distribuição pelos cromossomas destes SNPs novos comuns aos quatro genomas, observa-

se um maior número no cromossoma Y, ressaltando a ideia anteriormente explanada de que o cromossoma Y detém capacidade de caracterizar populacionalmente um individuo (Figura 45). Dos SNPs comuns, presentes em regiões codificantes, cerca de 5.004 foram SNPs não sinónimos (Tabela 11), e destes foram selecionados os SNPs cuja anotação pelo SIFT, PolyPhen, LRT e *Mutation Taster,* classificaram as variantes como deletérias. O conjunto de variantes previstas como deletérias em simultâneo pelos quatro *softwares* foi analisado no sentido de se encontrar o conjunto de genes onde estas variantes se localizavam. Obtiveram-se 11 genes que foram analisados relativamente às interações das proteínas que estes codificam, descritas na IntAct e representadas pelo Cytoscape (Anexo 6). Da análise do conjunto de genes referido, ressaltou a interação indireta entre o produto de quatro genes, *CDC27*, *CLIP1*, *VDR* e *PABPC1* (Figura 47). O gene *VDR*, que codifica o recetor da vitamina D, está alterado nos quatro genomas e apresenta uma destacada importância pela relação que estabelece com outros genes, também eles alterados nos quatro genomas, e que podem coadjuvar o potencial impacto deletério no metabolismo da vitamina D. O recetor da vitamina D demonstrou estar em interação, nomeadamente com a proteína do gene *RXRA*, associada à modulação da resposta à vitamina D (184). Esta proteína demonstrou interação, por sua vez, com a proteína codificada pelo gene *CDC27*, que contém uma alteração comum aos quatro genomas. A proteína expressa pelo gene *SNW1* é um cofator ativador da transcrição de vitamina D (185) e apresentou interação com a proteína do gene *VDR* e com a proteína expressa pelo gene *THRAP3*, descrita como associada ao recetor da vitamina D (GO:0042809). A proteína do gene *THRAP3*, por sua vez, apresentou interação com as proteínas dos genes identificados com variantes nos quatro genomas, *CLIP1* e *PABPC1* (Figura 47). A variante (rs2228570), identificada no gene *VDR* dos quatro genomas, está também associada à modulação da metabolização dos fármacos 1,25-dihidroxivitamina d3, calcipotriol, calcitriol e dexametasona, utilizados no tratamento de neoplasias da mama, fraturas ósseas, osteonecrose, neoplasias da próstata e na tuberculose (186-190). A relação que este farmacogene estabelece com a ocorrência de tuberculose, nomeadamente pelo impacto no metabolismo da vitamina D, mereceu a nossa especial atenção no que concerne ao estudo das características genéticas populacionais.

A tuberculose é uma doença infeciosa, causada pela infeção do bacilo *Mycobacterium tuberculosis (*bacilo de *Koch),* que é a principal causa da taxa de mortalidade por doenças infeciosas curáveis, tendo sido considera uma emergência global

pela Organização Mundial de Saúde (191). Em Portugal a taxa de incidência da tuberculose tem um nível intermédio, cerca de 21,6/100.000 habitantes, tendo vindo a reduzir nos últimos anos, embora Portugal apresente a taxa de incidência menos favorável de todos os países da europa ocidental (192). A infeção pelo vírus da imunodeficiência humana adquirida é o principal fator de risco para a ocorrência de tuberculose, estimando-se que a probabilidade aumente entre 21 a 34 vezes relativamente a um individuo saudável (193). Não obstante da influência deste fator, poderá estar em causa a possível ocorrência de uma suscetibilidade genética para a doença na população portuguesa. Têm sido publicados estudos nos últimos anos, que associam o défice da vitamina D ao aumento da ocorrência de tuberculose (194-196). Esta suscetibilidade advém do facto do metabolismo da vitamina D ter um papel preponderante ao nível da imunidade inata, nomeadamente pela ativação de macrófagos, restringindo o crescimento intracelular do bacilo de *Koch* (196). A ocorrência de alteração genética ao nível do gene *VDR* poderá revelar assim ser um fator de risco para a tuberculose. Dada a evidência destas variantes nos quatro genomas em estudo, e atendendo à taxa de incidência em Portugal ainda se encontrar acima da média de outros países europeus, nomeadamente em países com situações socio económicas semelhantes à portuguesa, realçamos esta associação como importante alvo de estudo futuro na população portuguesa.

A análise de todas as variantes não sinónimas, selecionadas a partir da análise da região codificante de cada genoma e passíveis de serem deletérias, revelou-se significativa na descoberta de possíveis suscetibilidades individuais e mesmo populacionais. Com a evolução do conhecimento no campo da genómica surgiu, contudo, a necessidade de caracterizar variantes presentes em regiões genómicas não codificantes e que, hoje se sabe, desempenham um papel importante na expressão génica. O mesmo raciocínio se tem aplicado às variantes sinónimas presentes nas regiões codificantes e que são excluídas da anotação funcional. Se analisarmos em detalhe as variantes exónicas identificadas no nosso estudo, relativamente à anotação funcional realizada em cada genoma, deparamo-nos com valores próximos de SNPs sinónimos e não sinónimos (Tabelas 5 e 6). De igual modo, verificou-se nos quatro genomas, uma proximidade entre o número de INDELs *frameshift* e não *frameshift* (Tabelas 7 e 8). Evidências recentes têm revelado que existem associações a fenótipos deletérios, quer em variantes não sinónimas quer em variantes sinónimas, em regiões codificantes e não codificantes (197). Chen e colaboradores encontraram um

número semelhante de SNPs sinónimos e não sinónimos que partilham o mesmo *odds ratio* de associação a doença e com uma dimensão do efeito semelhante (197). Mais, este estudo observou valores semelhantes também para variantes presentes nas regiões 5'UTR e 3'UTR, podendo as últimas estar relacionadas com o processo regulador de miRNAs e, portanto, a exclusão do seu estudo poderá ser uma causa da perda de herdabilidade verificada em estudos de associação genótipo-fenótipo.

No seguimento deste entendimento, o nosso estudo abrangeu a análise de elementos não codificantes, os miRNAs, tendo sido relacionados com a expressão dos genes que contêm variantes exónicas em cada genoma (Figuras 50, 52, 54 e 56). Embora esta análise agregadora tenha permitido realizar uma nova abordagem, outros elementos reguladores, que se acredita também contribuam de forma substancial para o funcionamento do organismo humano, ainda carecem de análise.

Algumas descobertas recentes têm atribuído um importante papel aos rSNPs, que se encontram em regiões não codificantes, nomeadamente em mecanismos moleculares de doenças complexas (198). Pela primeira vez, foi identificada a percentagem de SNPs intrónicos correspondente a rSNPs, e a SNPs em LD com rSNPs, em quatro genomas completos (Tabela 12). Dos SNPs intrónicos identificados nos quatro genomas, ≈80% corresponderam a SNPs cujo efeito regulador da expressão génica encontra-se descrito em bases de dados. A elevada fração de rSNPs intrónicos obtida em cada genoma, reforça a necessidade de estudar as variações presentes em regiões não codificantes, potencial forma de perda de herdabilidade na análise de fenótipos complexos centrada na região codificante, anteriormente discutida. Um estudo de GWAS publicado, havia já reportado um conjunto significativo de SNPs implicados na associação a doenças complexas e localizados na região intergénica e intrónica (198), indicando que muitos SNPs considerados de risco podem afetar fenótipos de uma forma não codificante, nomeadamente através do seu impacto na regulação génica. Dos rSNPs presentes na região intrónica identificados nos quatro genomas, ≈20% remetiam para um tipo de regulação transcricional proximal e ≈31% para um tipo de regulação distal (Tabela 12). Estes resultados realçam a importância de compreender a molécula de DNA na sua tridimensionalidade e a forma como a proximidade não só na sequência linear, mas espacial, influencia a própria expressão génica. Dos rSNPs intrónicos identificados 91%

em cada genoma apresentaram também associação à regulação pós-transcricional mediada por proteína de ligação ao RNA. Os valores obtidos poderão ser explicados pelo elevado número de proteínas que se relacionam com o pré-mRNA para regular o seu processamento, nomeadamente ao nível da edição, *splicing* e poliadenilação, e estabilidade do mRNA, nomeadamente, tornando o complexo ribonucleoproteico que se forma entre estas proteínas e o seu mRNA alvo, praticamente único (199). Os valores obtidos na identificação de rSNPs com diferentes tipos de regulação evidenciam a necessidade de introduzir a análise destes SNPs em estudos futuros, nomeadamente em estudos direcionados para determinada condição ou doença.

A necessidade expressa de analisar as alterações presentes num determinado indivíduo de modo a identificar associações a determinados fenótipos, sairia também beneficiada com a inclusão de características populacionais nessa análise. A diversidade genética é intimamente relacionada com a ancestralidade genética populacional, o que torna relevante o estudo da ancestralidade dos quatro genomas de indivíduos portugueses em análise.

Compreender a ancestralidade genética de um determinado individuo passa por compreender o processo inicial de migração populacional, que resultou na distribuição de populações que hoje conhecemos, bem como o fluxo génico entre populações que ao longo da evolução temporal foi ocorrendo. A migração de populações, a partir de uma população parental, pensa-se que terá tido início acerca de 100 mil anos atrás, de uma pequena população de 1.000 indivíduos (uma tribo), muito provavelmente no leste de África (200). Dados paleoantropológicos e paleoclimatológicos têm permitido inferir o movimento migratório a partir de África e sustentado o modelo da evolução humana moderna, também chamado de *Out-of-Africa 2* (200-202). Segundo este modelo, a pequena população africana expandiu-se pelo restante continente africano, tendo ocorrido posteriormente, acerca de 60 mil a 40 mil anos atrás, a segunda expansão de uma população descendente para a Ásia e daí para os restantes continentes. Informações genéticas dão indicações de que a expansão dos humanos modernos para a Ásia ocorreu por duas rotas. A primeira, a rota do sul, terá ocorrido ao longo da costa sul e ocidental da Ásia, a partir da qual se bifurcou para o norte e sul (203). No sul, estes humanos modernos alcançaram a Oceânia, enquanto a expansão do norte só mais tarde alcançou a China, o Japão e a América, acerca

de 35 mil a 15 mil anos. A segunda foi a rota central, de onde a migração progrediu em todas as direções, encontrando a Europa e a Ásia do norte e oriental, acerca de 40 mil anos, depois da qual a primeira migração para a américa ocorreu (ainda antes da descrita na rota do sul) (204). Este é o modelo de migração de populações mais aceite, continuando apenas por esclarecer se a divergência entre as duas rotas ocorreu em África, ou depois de entrarem na Ásia ocidental. Tendo por base este modelo, analisamos o resultado obtido após a PCA relativamente à ancestralidade de cada um dos quatro genomas portugueses (Figura 57).

Realçamos, primeiramente, a proximidade que existe no posicionamento dos quatro genomas portugueses analisados, na representação da PCA (Figuras 57 e 58). Tal facto poderá dar indicação da adequação da metodologia utilizada na análise, bem como da qualidade da mesma. De acordo com o esperado, foi também a localização dos quatro genomas portugueses no bloco em que se agruparam as populações europeias estudadas.

Foram observados 9 blocos de populações, que se agruparam pela proximidade genética verificada (Figura 57). A distância verificada no bloco que engloba a maioria das populações africanas introduzidas no estudo, dos restantes blocos que englobam as populações dos restantes continentes, fornece a visão global da migração humana inicial, *Out-of-Africa 2*, referida anteriormente (202). O bloco das populações europeias, no qual se incluem os genomas portugueses em estudo, mostrou um pequeno isolamento relativamente às populações asiáticas ocidentais, do centro e do sul, bem como do norte de África. Tal evidência reflete também o elevado fluxo génico ao longo dos últimos 80 mil anos, com a migração das populações e inúmeras expansões entre estas populações (205). Relativamente às populações europeias, os quatro genomas portugueses encontram-se geneticamente próximos das populações italianas, francesa e de uma população da Ásia ocidental (Italian, Tuscan, French e Druze). A proximidade genética evidenciada pelos indivíduos portugueses à população asiática (Druze), superior à proximidade que os indivíduos evidenciaram relativamente à população do norte de África (Mozabite), reflete provavelmente os efeitos da migração das populações, corroborando a ideia de que a entrada das populações na Europa não ocorreu pelo norte de África, mas sim pelo continente asiático (Figura 57). A proximidade que também se observou entre estes indivíduos e as populações francesa (French) e italianas (Italian, Tuscan) poderá ser

resultado do elevado fluxo génico entre as populações europeias (Figura 58). As populações Russas (Russian e Adygei) estão mais afastadas dos indivíduos portugueses bem como das restantes populações europeias, denotando uma clara referência asiática. As populações do País Basco e da Sardenha (Basque; Sardainian) são as mais distantes das populações asiáticas, o que pode refletir uma menor influência da migração humana inicial.

A análise global realizada às variações genéticas identificadas nos genomas dos quatro indivíduos portugueses oferece uma poderosa oportunidade de compreender a expansão e migração de populações, que faculta a estes genomas a sua ancestralidade genética.

Conclusão

A análise dos primeiros quatro genomas de indivíduos portugueses permitiu caracterizar individualmente a estrutura genómica em termos de variação existente. Em cada genoma foi verificado que, em média, 11% dos SNPs e 28,9% dos INDELs foram identificados pela primeira vez através deste estudo. As variantes identificadas encontraram-se numa proporção heterozigotia/homozigotia de 2/1, tendo sido verificado que o cromossoma 21 apresentou a maior densidade de alterações nos quatro genomas portugueses. O estudo desenvolveu uma análise integrada da variação com maior impacto funcional previsto, nomeadamente a variação exónica não sinónima, a variação presente em miRNAs e a variação que modula a resposta terapêutica. A determinação da ancestralidade genética dos indivíduos portugueses permitiu diferenciá-los em termos populacionais, posicionando-os próximo de outras populações caucasianas. A análise global da variabilidade genética da população poderá contribuir para a identificação de suscetibilidades a doenças, o que se torna preponderante na atuação preventiva que sustenta a preservação e evolução de cada indivíduo da população, e que é a principal meta da medicina personalizada.

Referências

1. Watson JD, Crick FH. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature. 1953;171(4356):737-8.
2. Regateiro FJ. Manual de genética médica. 1ª ed., 2ª reimpressão (2007) ed. Coimbra: Imprensa da Universidade de Coimbra; 2003 2003. 514 p.
3. Gilbert W, Maxam A. The nucleotide sequence of the lac operator. Proceedings of the National Academy of Sciences of the United States of America. 1973;70(12):3581-4.
4. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America. 1977;74(12):5463-7.
5. Sears LE, Moran LS, Kissinger C, Creasey T, Perry-O'Keefe H, Roskey M, Sutherland E, Slatko BE. CircumVent thermal cycle sequencing and alternative manual and automated DNA sequencing protocols using the highly thermostable VentR (exo-) DNA polymerase. BioTechniques. 1992;13(4):626-33.
6. McBride LJ, Koepf SM, Gibbs RA, Salser W, Mayrand PE, Hunkapiller MW, Kronick MN. Automated DNA sequencing methods involving polymerase chain reaction. Clinical chemistry. 1989;35(11):2196-201.
7. Panussis DA, Cook MW, Rifkin LL, Snider JE, Strong JT, McGrane RM, Wilson RK, Mardis ER. A pneumatic device for rapid loading of DNA sequencing gels. Genome research. 1998;8(5):543-8.
8. Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome research. 1998;8(3):186-94.
9. Wong LJC. Next Generation Sequencing: Translation to Clinical Diagnostics2013.
10. Research NCfHG, Health USDoEOo, Program ERHG. Understanding our genetic inheritance: the U.S. Human Genome Project : the first five years, FY 1991-1995: U.S. Dept. of Health and Human Services, Public Health Service, National Institutes of Health, National Center for Human Genome Research; 1990.
11. Collins F, Galas D. A new five-year plan for the U.S. Human Genome Project. Science. 1993;262(5130):43-6.
12. Venter JC, Adams MD, Sutton GG, Kerlavage AR, Smith HO, Hunkapiller M. Shotgun sequencing of the human genome. Science. 1998;280(5369):1540-2.
13. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs

RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860-921.

14. Finishing the euchromatic sequence of the human genome. Nature. 2004;431(7011):931-45.
15. Eisenhaber F. A decade after the first full human genome sequencing: when will we understand our own genome? Journal of bioinformatics and computational biology. 2012;10(5):1271001.
16. Mardis ER. Next-generation sequencing platforms. Annu Rev Anal Chem (Palo Alto Calif). 2013;6:287-303.
17. Shendure J, Ji H. Next-generation DNA sequencing. Nature biotechnology. 2008;26(10):1135-45.
18. Mardis ER. Next-generation DNA sequencing methods. Annual review of genomics and human genetics. 2008;9:387-402.
19. Bonetta L. Whole-genome sequencing breaks the cost barrier. Cell. 2010;141(6):917-9.
20. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome research. 2008;18(11):1851-8.
21. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010;26(5):589-95.
22. Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S, Berlin AM, Aird D, Costello M, Daza R, Williams L, Nicol R, Gnirke A, Nusbaum C, Lander ES, Jaffe DB. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proceedings of the National Academy of Sciences of the United States of America. 2011;108(4):1513-8.
23. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan

MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005;437(7057):376-80.
24. Dressman D, Yan H, Traverso G, Kinzler KW, Vogelstein B. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. Proceedings of the National Academy of Sciences of the United States of America. 2003;100(15):8817-22.
25. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM. Accurate multiplex polony sequencing of an evolved bacterial genome. Science. 2005;309(5741):1728-32.
26. Fedurco M, Romieu A, Williams S, Lawrence I, Turcatti G. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. Nucleic acids research. 2006;34(3):e22.
27. Adessi C, Matton G, Ayala G, Turcatti G, Mermod JJ, Mayer P, Kawashima E. Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. Nucleic acids research. 2000;28(20):E87.
28. Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, Leamon JH, Johnson K, Milgrew MJ, Edwards M, Hoon J, Simons JF, Marran D, Myers JW, Davidson JF, Branting A, Nobile JR, Puc BP, Light D, Clark TA, Huber M, Branciforte JT, Stoner IB, Cawley SE, Lyons M, Fu Y, Homer N, Sedova M, Miao X, Reed B, Sabina J, Feierstein E, Schorn M, Alanjary M, Dimalanta E, Dressman D, Kasinskas R, Sokolsky T, Fidanza JA, Namsaraev E, McKernan KJ, Williams A, Roth GT, Bustillo J. An integrated semiconductor device enabling non-optical genome sequencing. Nature. 2011;475(7356):348-52.
29. Milos PM. Emergence of single-molecule sequencing and potential for molecular diagnostic applications. Expert review of molecular diagnostics. 2009;9(7):659-66.
30. Clarke J, Wu HC, Jayasinghe L, Patel A, Reid S, Bayley H. Continuous base identification for single-molecule nanopore DNA sequencing. Nature nanotechnology. 2009;4(4):265-70.
31. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, Dewinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, Lundquist P, Ma C, Marks P, Maxham M, Murphy D, Park I, Pham T, Phillips M, Roy J, Sebra R, Shen G, Sorenson J, Tomaney A, Travers K, Trulson M, Vieceli J, Wegener J, Wu D, Yang A, Zaccarin D, Zhao P, Zhong F, Korlach J, Turner S. Real-time DNA sequencing from single polymerase molecules. Science. 2009;323(5910):133-8.
32. Church GM. The personal genome project. Molecular systems biology. 2005;1:2005 0030.
33. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, Lin Y, MacDonald JR, Pang AW, Shago M, Stockwell TB, Tsiamouri A, Bafna V, Bansal V, Kravitz SA, Busam DA, Beeson KY, McIntosh TC, Remington KA, Abril JF, Gill J, Borman J, Rogers YH, Frazier ME, Scherer SW, Strausberg RL, Venter JC. The diploid genome sequence of an individual human. PLoS biology. 2007;5(10):e254.
34. Ng PC, Kirkness EF. Whole genome sequencing. Methods Mol Biol. 2010;628:215-26.
35. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Lupski JR, Chinault C, Song XZ, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Margulies M,

Weinstock GM, Gibbs RA, Rothberg JM. The complete genome of an individual by massively parallel DNA sequencing. Nature. 2008;452(7189):872-6.

36. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, Rasolonjatovo IM, Reed MT, Rigatti R, Rodighiero C, Ross MT, Sabot A, Sankar SV, Scally A, Schroth GP, Smith ME, Smith VP, Spiridou A, Torrance PE, Tzonev SS, Vermaas EH, Walter K, Wu X, Zhang L, Alam MD, Anastasi C, Aniebo IC, Bailey DM, Bancarz IR, Banerjee S, Barbour SG, Baybayan PA, Benoit VA, Benson KF, Bevis C, Black PJ, Boodhun A, Brennan JS, Bridgham JA, Brown RC, Brown AA, Buermann DH, Bundu AA, Burrows JC, Carter NP, Castillo N, Chiara ECM, Chang S, Neil Cooley R, Crake NR, Dada OO, Diakoumakos KD, Dominguez-Fernandez B, Earnshaw DJ, Egbujor UC, Elmore DW, Etchin SS, Ewan MR, Fedurco M, Fraser LJ, Fuentes Fajardo KV, Scott Furey W, George D, Gietzen KJ, Goddard CP, Golda GS, Granieri PA, Green DE, Gustafson DL, Hansen NF, Harnish K, Haudenschild CD, Heyer NI, Hims MM, Ho JT, Horgan AM, Hoschler K, Hurwitz S, Ivanov DV, Johnson MQ, James T, Huw Jones TA, Kang GD, Kerelska TH, Kersey AD, Khrebtukova I, Kindwall AP, Kingsbury Z, Kokko-Gonzales PI, Kumar A, Laurent MA, Lawley CT, Lee SE, Lee X, Liao AK, Loch JA, Lok M, Luo S, Mammen RM, Martin JW, McCauley PG, McNitt P, Mehta P, Moon KW, Mullens JW, Newington T, Ning Z, Ling Ng B, Novo SM, O'Neill MJ, Osborne MA, Osnowski A, Ostadan O, Paraschos LL, Pickering L, Pike AC, Chris Pinkard D, Pliskin DP, Podhasky J, Quijano VJ, Raczy C, Rae VH, Rawlings SR, Chiva Rodriguez A, Roe PM, Rogers J, Rogert Bacigalupo MC, Romanov N, Romieu A, Roth RK, Rourke NJ, Ruediger ST, Rusman E, Sanches-Kuiper RM, Schenker MR, Seoane JM, Shaw RJ, Shiver MK, Short SW, Sizto NL, Sluis JP, Smith MA, Ernest Sohna Sohna J, Spence EJ, Stevens K, Sutton N, Szajkowski L, Tregidgo CL, Turcatti G, Vandevondele S, Verhovsky Y, Virk SM, Wakelin S, Walcott GC, Wang J, Worsley GJ, Yan J, Yau L, Zuerlein M, Mullikin JC, Hurles ME, McCooke NJ, West JS, Oaks FL, Lundberg PL, Klenerman D, Durbin R, Smith AJ. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008;456(7218):53-9.
37. Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J, Guo Y, Feng B, Li H, Lu Y, Fang X, Liang H, Du Z, Li D, Zhao Y, Hu Y, Yang Z, Zheng H, Hellmann I, Inouye M, Pool J, Yi X, Zhao J, Duan J, Zhou Y, Qin J, Ma L, Li G, Zhang G, Yang B, Yu C, Liang F, Li W, Li S, Ni P, Ruan J, Li Q, Zhu H, Liu D, Lu Z, Li N, Guo G, Ye J, Fang L, Hao Q, Chen Q, Liang Y, Su Y, San A, Ping C, Yang S, Chen F, Li L, Zhou K, Ren Y, Yang L, Gao Y, Yang G, Li Z, Feng X, Kristiansen K, Wong GK, Nielsen R, Durbin R, Bolund L, Zhang X, Yang H. The diploid genome sequence of an Asian individual. Nature. 2008;456(7218):60-5.
38. Ahn SM, Kim TH, Lee S, Kim D, Ghang H, Kim DS, Kim BC, Kim SY, Kim WY, Kim C, Park D, Lee YS, Kim S, Reja R, Jho S, Kim CG, Cha JY, Kim KH, Lee B, Bhak J, Kim SJ. The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. Genome research. 2009;19(9):1622-9.
39. Kim JI, Ju YS, Park H, Kim S, Lee S, Yi JH, Mudge J, Miller NA, Hong D, Bell CJ, Kim HS, Chung IS, Lee WC, Lee JS, Seo SH, Yun JY, Woo HN, Lee H, Suh D, Kim HJ, Yavartanoo M, Kwak M, Zheng Y, Lee MK, Kim JY, Gokcumen O, Mills RE, Zaranek AW, Thakuria J, Wu X, Kim RW, Huntley JJ, Luo S, Schroth GP, Wu TD, Kim H, Yang KS, Park WY,

Church GM, Lee C, Kingsmore SF, Seo JS. A highly annotated whole-genome sequence of a Korean individual. Nature. 2009;460(7258):1011-5.
40. Fujimoto A, Nakagawa H, Hosono N, Nakano K, Abe T, Boroevich KA, Nagasaki M, Yamaguchi R, Shibuya T, Kubo M, Miyano S, Nakamura Y, Tsunoda T. Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massively parallel sequencing. Nature genetics. 2010;42(11):931-6.
41. Tong P, Prendergast JG, Lohan AJ, Farrington SM, Cronin S, Friel N, Bradley DG, Hardiman O, Evans A, Wilson JF, Loftus B. Sequencing and analysis of an Irish human genome. Genome Biol. 2010;11(9):R91.
42. Gupta R, Ratan A, Rajesh C, Chen R, Kim HL, Burhans R, Miller W, Santhosh S, Davuluri RV, Butte AJ, Schuster SC, Seshagiri S, Thomas G. Sequencing and analysis of a South Asian-Indian personal genome. BMC genomics. 2012;13:440.
43. Dogan H, Can H, Otu HH. Whole genome sequence of a Turkish individual. PloS one. 2014;9(1):e85233.
44. Ginsburg GS, Willard HF. Genomic and personalized medicine. 2nd ed. London ; Waltham, MA: Elsevier/Academic Press; 2013.
45. Gonzaga-Jauregui C, Lupski JR, Gibbs RA. Human genome sequencing in health and disease. Annual review of medicine. 2012;63:35-61.
46. Garrod AE. About Alkaptonuria. Medico-chirurgical transactions. 1902;85:69-78.
47. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56-65.
48. Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, Bonnen PE, de Bakker PI, Deloukas P, Gabriel SB, Gwilliam R, Hunt S, Inouye M, Jia X, Palotie A, Parkin M, Whittaker P, Chang K, Hawes A, Lewis LR, Ren Y, Wheeler D, Muzny DM, Barnes C, Darvishi K, Hurles M, Korn JM, Kristiansson K, Lee C, McCarrol SA, Nemesh J, Keinan A, Montgomery SB, Pollack S, Price AL, Soranzo N, Gonzaga-Jauregui C, Anttila V, Brodeur W, Daly MJ, Leslie S, McVean G, Moutsianas L, Nguyen H, Zhang Q, Ghori MJ, McGinnis R, McLaren W, Takeuchi F, Grossman SR, Shlyakhter I, Hostetter EB, Sabeti PC, Adebamowo CA, Foster MW, Gordon DR, Licinio J, Manca MC, Marshall PA, Matsuda I, Ngare D, Wang VO, Reddy D, Rotimi CN, Royal CD, Sharp RR, Zeng C, Brooks LD, McEwen JE. Integrating common and rare genetic variation in diverse human populations. Nature. 2010;467(7311):52-8.
49. Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. A map of human genome variation from population-scale sequencing. Nature. 2010;467(7319):1061-73.
50. The International HapMap Project. Nature. 2003;426(6968):789-96.
51. Manolio TA, Collins FS. The HapMap and genome-wide association studies in diagnosis and therapy. Annual review of medicine. 2009;60:443-56.
52. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM. Finding the missing heritability of complex diseases. Nature. 2009;461(7265):747-53.
53. Siva N. 1000 Genomes project. Nature biotechnology. 2008;26(3):256.

54. Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, Gabriel S, Rieder MJ, Altshuler D, Shendure J, Nickerson DA, Bamshad MJ, Akey JM. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature. 2013;493(7431):216-20.
55. Boi L. Epigenetic phenomena, chromatin dynamics, and gene expression. New theoretical approaches in the study of living systems. Rivista di biologia. 2008;101(3):405-42.
56. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science. 2004;306(5696):636-40.
57. Qu H, Fang X. A brief review on the Human Encyclopedia of DNA Elements (ENCODE) project. Genomics, proteomics & bioinformatics. 2013;11(3):135-41.
58. Sboner A, Mu XJ, Greenbaum D, Auerbach RK, Gerstein MB. The real cost of sequencing: higher than you think! Genome biology. 2011;12(8):125.
59. Schadt EE, Linderman MD, Sorenson J, Lee L, Nolan GP. Computational solutions to large-scale data management and analysis. Nature reviews Genetics. 2010;11(9):647-57.
60. Pabinger S, Dander A, Fischer M, Snajder R, Sperk M, Efremova M, Krabichler B, Speicher MR, Zschocke J, Trajanoski Z. A survey of tools for variant analysis of next-generation genome sequencing data. Briefings in bioinformatics. 2013.
61. Dai M, Thompson RC, Maher C, Contreras-Galindo R, Kaplan MH, Markovitz DM, Omenn G, Meng F. NGSQC: cross-platform quality analysis pipeline for deep sequencing data. BMC genomics. 2010;11 Suppl 4:S7.
62. Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome research. 1998;8(3):175-85.
63. Pavlopoulos GA, Oulas A, Iacucci E, Sifrim A, Moreau Y, Schneider R, Aerts J, Iliopoulos I. Unraveling genomic variation from next generation sequencing data. BioData mining. 2013;6(1):13.
64. Bromberg Y. Building a genome analysis pipeline to predict disease risk and prevent disease. Journal of molecular biology. 2013;425(21):3993-4005.
65. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic acids research. 2010;38(16):e164.
66. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. Nucleic acids research. 2001;29(1):308-11.
67. Jorde LB, Watkins WS, Bamshad MJ. Population genomics: a bridge from evolutionary history to genetic medicine. Human molecular genetics. 2001;10(20):2199-207.
68. Rotimi CN, Jorde LB. Ancestry and disease in the age of genomic medicine. The New England journal of medicine. 2010;363(16):1551-8.
69. Nei M. F-statistics and analysis of gene diversity in subdivided populations. Annals of human genetics. 1977;41(2):225-33.
70. Jorde LB, Watkins WS, Bamshad MJ, Dixon ME, Ricker CE, Seielstad MT, Batzer MA. The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data. American journal of human genetics. 2000;66(3):979-88.
71. Matullo G, Di Gaetano C, Guarrera S. Next generation sequencing and rare genetic variants: from human population studies to medical genetics. Environmental and molecular mutagenesis. 2013;54(7):518-32.
72. Nei M. Analysis of gene diversity in subdivided populations. Proceedings of the National Academy of Sciences of the United States of America. 1973;70(12):3321-3.
73. Watkins WS, Rogers AR, Ostler CT, Wooding S, Bamshad MJ, Brassington AM, Carroll ML, Nguyen SV, Walker JA, Prasad BV, Reddy PG, Das PK, Batzer MA, Jorde LB. Genetic

variation among world populations: inferences from 100 Alu insertion polymorphisms. Genome research. 2003;13(7):1607-18.
74. Bamshad MJ, Olson SE. Does race exist? Scientific American. 2003;289(6):78-85.
75. Burnett MS, Strain KJ, Lesnick TG, de Andrade M, Rocca WA, Maraganore DM. Reliability of self-reported ancestry among siblings: implications for genetic association studies. American journal of epidemiology. 2006;163(5):486-92.
76. Liu Y, Nyunoya T, Leng S, Belinsky SA, Tesfaigzi Y, Bruse S. Softwares and methods for estimating genetic ancestry in human populations. Human genomics. 2013;7:1.
77. Wang C, Zhan X, Bragg-Gresham J, Kang HM, Stambolian D, Chew EY, Branham KE, Heckenlively J, Fulton R, Wilson RK, Mardis ER, Lin X, Swaroop A, Zollner S, Abecasis GR. Ancestry estimation and control of population stratification for sequence-based association studies. Nature genetics. 2014;46(4):409-15.
78. Wang C, Zollner S, Rosenberg NA. A quantitative comparison of the similarity between genes and geography in worldwide human populations. PLoS genetics. 2012;8(8):e1002886.
79. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Federhen S, Feolo M, Fingerman IM, Geer LY, Helmberg W, Kapustin Y, Landsman D, Lipman DJ, Lu Z, Madden TL, Madej T, Maglott DR, Marchler-Bauer A, Miller V, Mizrachi I, Ostell J, Panchenko A, Phan L, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Shumway M, Sirotkin K, Slotta D, Souvorov A, Starchenko G, Tatusova TA, Wagner L, Wang Y, Wilbur WJ, Yaschenko E, Ye J. Database resources of the National Center for Biotechnology Information. Nucleic acids research. 2011;39(Database issue):D38-51.
80. Thorn CF, Klein TE, Altman RB. PharmGKB: the Pharmacogenomics Knowledge Base. Methods Mol Biol. 2013;1015:311-20.
81. Ashley EA, Butte AJ, Wheeler MT, Chen R, Klein TE, Dewey FE, Dudley JT, Ormond KE, Pavlovic A, Morgan AA, Pushkarev D, Neff NF, Hudgins L, Gong L, Hodges LM, Berlin DS, Thorn CF, Sangkuhl K, Hebert JM, Woon M, Sagreiya H, Whaley R, Knowles JW, Chou MF, Thakuria JV, Rosenbaum AM, Zaranek AW, Church GM, Greely HT, Quake SR, Altman RB. Clinical assessment incorporating a personal genome. Lancet. 2010;375(9725):1525-35.
82. Badano JL, Katsanis N. Beyond Mendel: an evolving view of human genetic disease transmission. Nature reviews Genetics. 2002;3(10):779-89.
83. Lupski JR, Reid JG, Gonzaga-Jauregui C, Rio Deiros D, Chen DC, Nazareth L, Bainbridge M, Dinh H, Jing C, Wheeler DA, McGuire AL, Zhang F, Stankiewicz P, Halperin JJ, Yang C, Gehman C, Guo D, Irikat RK, Tom W, Fantin NJ, Muzny DM, Gibbs RA. Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. The New England journal of medicine. 2010;362(13):1181-91.
84. Baranzini SE, Mudge J, van Velkinburgh JC, Khankhanian P, Khrebtukova I, Miller NA, Zhang L, Farmer AD, Bell CJ, Kim RW, May GD, Woodward JE, Caillier SJ, McElroy JP, Gomez R, Pando MJ, Clendenen LE, Ganusova EE, Schilkey FD, Ramaraj T, Khan OA, Huntley JJ, Luo S, Kwok PY, Wu TD, Schroth GP, Oksenberg JR, Hauser SL, Kingsmore SF. Genome, epigenome and RNA sequences of monozygotic twins discordant for multiple sclerosis. Nature. 2010;464(7293):1351-6.
85. Day-Williams AG, Zeggini E. The effect of next-generation sequencing technology on complex trait research. European journal of clinical investigation. 2011;41(5):561-7.
86. Marian AJ. Molecular genetic studies of complex phenotypes. Translational research : the journal of laboratory and clinical medicine. 2012;159(2):64-79.

87. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. American journal of human genetics. 2012;90(1):7-24.
88. Vilarino-Guell C, Wider C, Ross OA, Dachsel JC, Kachergus JM, Lincoln SJ, Soto-Ortolaza AI, Cobb SA, Wilhoite GJ, Bacon JA, Behrouz B, Melrose HL, Hentati E, Puschmann A, Evans DM, Conibear E, Wasserman WW, Aasly JO, Burkhard PR, Djaldetti R, Ghika J, Hentati F, Krygowska-Wajs A, Lynch T, Melamed E, Rajput A, Rajput AH, Solida A, Wu RM, Uitti RJ, Wszolek ZK, Vingerhoets F, Farrer MJ. VPS35 mutations in Parkinson disease. American journal of human genetics. 2011;89(1):162-7.
89. Kathiresan S, Srivastava D. Genetics of human cardiovascular disease. Cell. 2012;148(6):1242-57.
90. Erlich HA, Valdes AM, McDevitt SL, Simen BB, Blake LA, McGowan KR, Todd JA, Rich SS, Noble JA. Next generation sequencing reveals the association of DRB3*02:02 with type 1 diabetes. Diabetes. 2013;62(7):2618-22.
91. Berger MF, Lawrence MS, Demichelis F, Drier Y, Cibulskis K, Sivachenko AY, Sboner A, Esgueva R, Pflueger D, Sougnez C, Onofrio R, Carter SL, Park K, Habegger L, Ambrogio L, Fennell T, Parkin M, Saksena G, Voet D, Ramos AH, Pugh TJ, Wilkinson J, Fisher S, Winckler W, Mahan S, Ardlie K, Baldwin J, Simons JW, Kitabayashi N, MacDonald TY, Kantoff PW, Chin L, Gabriel SB, Gerstein MB, Golub TR, Meyerson M, Tewari A, Lander ES, Getz G, Rubin MA, Garraway LA. The genomic complexity of primary human prostate cancer. Nature. 2011;470(7333):214-20.
92. Mardis ER, Wilson RK. Cancer genome sequencing: a review. Human molecular genetics. 2009;18(R2):R163-8.
93. Castle JC, Kreiter S, Diekmann J, Lower M, van de Roemer N, de Graaf J, Selmi A, Diken M, Boegel S, Paret C, Koslowski M, Kuhn AN, Britten CM, Huber C, Tureci O, Sahin U. Exploiting the mutanome for tumor vaccination. Cancer research. 2012;72(5):1081-91.
94. Clark MB, Amaral PP, Schlesinger FJ, Dinger ME, Taft RJ, Rinn JL, Ponting CP, Stadler PF, Morris KV, Morillon A, Rozowsky JS, Gerstein MB, Wahlestedt C, Hayashizaki Y, Carninci P, Gingeras TR, Mattick JS. The reality of pervasive transcription. PLoS biology. 2011;9(7):e1000625; discussion e1102.
95. Lee RC, Feinbaum RL, Ambros V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell. 1993;75(5):843-54.
96. Derrien T, Guigo R, Johnson R. The Long Non-Coding RNAs: A New (P)layer in the "Dark Matter". Frontiers in genetics. 2011;2:107.
97. Moazed D. Small RNAs in transcriptional gene silencing and genome defence. Nature. 2009;457(7228):413-20.
98. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, Lagarde J, Veeravalli L, Ruan X, Ruan Y, Lassmann T, Carninci P, Brown JB, Lipovich L, Gonzalez JM, Thomas M, Davis CA, Shiekhattar R, Gingeras TR, Hubbard TJ, Notredame C, Harrow J, Guigo R. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome research. 2012;22(9):1775-89.
99. Bartel DP. MicroRNAs: Target Recognition and Regulatory Functions. Cell. 2009;136(2):215-33.
100. Crick F. Central dogma of molecular biology. Nature. 1970;227(5258):561-3.

101. Reinhart BJ, Slack FJ, Basson M, Pasquinelli AE, Bettinger JC, Rougvie AE, Horvitz HR, Ruvkun G. The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. Nature. 2000;403(6772):901-6.
102. Dews M, Homayouni A, Yu D, Murphy D, Sevignani C, Wentzel E, Furth EE, Lee WM, Enders GH, Mendell JT, Thomas-Tikhonenko A. Augmentation of tumor angiogenesis by a Myc-activated microRNA cluster. Nature genetics. 2006;38(9):1060-5.
103. Kulshreshtha R, Ferracin M, Wojcik SE, Garzon R, Alder H, Agosto-Perez FJ, Davuluri R, Liu CG, Croce CM, Negrini M, Calin GA, Ivan M. A microRNA signature of hypoxia. Molecular and cellular biology. 2007;27(5):1859-67.
104. Zhou X, Ruan J, Wang G, Zhang W. Characterization and identification of microRNA core promoters in four model species. PLoS computational biology. 2007;3(3):e37.
105. Sun W, Julie Li YS, Huang HD, Shyy JY, Chien S. microRNA: a master regulator of cellular processes for bioengineering systems. Annual review of biomedical engineering. 2010;12:1-27.
106. Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic acids research. 2013.
107. Denli AM, Tops BB, Plasterk RH, Ketting RF, Hannon GJ. Processing of primary microRNAs by the Microprocessor complex. Nature. 2004;432(7014):231-5.
108. Yi R, Qin Y, Macara IG, Cullen BR. Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. Genes & development. 2003;17(24):3011-6.
109. Hutvagner G, McLachlan J, Pasquinelli AE, Balint E, Tuschl T, Zamore PD. A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. Science. 2001;293(5531):834-8.
110. Fabian MR, Sonenberg N, Filipowicz W. Regulation of mRNA translation and stability by microRNAs. Annual review of biochemistry. 2010;79:351-79.
111. Westholm JO, Lai EC. Mirtrons: microRNA biogenesis via splicing. Biochimie. 2011;93(11):1897-904.
112. Wang Y, Juranek S, Li H, Sheng G, Tuschl T, Patel DJ. Structure of an argonaute silencing complex with a seed-containing guide DNA and target RNA duplex. Nature. 2008;456(7224):921-6.
113. Grimson A, Farh KK, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Molecular cell. 2007;27(1):91-105.
114. Kiriakidou M, Tan GS, Lamprinaki S, De Planell-Saguer M, Nelson PT, Mourelatos Z. An mRNA m7G cap binding-like motif within human Ago2 represses translation. Cell. 2007;129(6):1141-51.
115. Petersen CP, Bordeleau ME, Pelletier J, Sharp PA. Short RNAs repress translation after initiation in mammalian cells. Molecular cell. 2006;21(4):533-42.
116. Bernstein E, Kim SY, Carmell MA, Murchison EP, Alcorn H, Li MZ, Mills AA, Elledge SJ, Anderson KV, Hannon GJ. Dicer is essential for mouse development. Nature genetics. 2003;35(3):215-7.
117. Erson AE, Petty EM. MicroRNAs in development and disease. Clinical genetics. 2008;74(4):296-306.
118. Poy MN, Eliasson L, Krutzfeldt J, Kuwajima S, Ma X, Macdonald PE, Pfeffer S, Tuschl T, Rajewsky N, Rorsman P, Stoffel M. A pancreatic islet-specific microRNA regulates insulin secretion. Nature. 2004;432(7014):226-30.

119. Mersey BD, Jin P, Danner DJ. Human microRNA (miR29b) expression controls the amount of branched chain alpha-ketoacid dehydrogenase complex in a cell. Human molecular genetics. 2005;14(22):3371-7.
120. Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, Sweet-Cordero A, Ebert BL, Mak RH, Ferrando AA, Downing JR, Jacks T, Horvitz HR, Golub TR. MicroRNA expression profiles classify human cancers. Nature. 2005;435(7043):834-8.
121. Meng F, Henson R, Wehbe-Janek H, Ghoshal K, Jacob ST, Patel T. MicroRNA-21 regulates expression of the PTEN tumor suppressor gene in human hepatocellular cancer. Gastroenterology. 2007;133(2):647-58.
122. Shilo S, Roy S, Khanna S, Sen CK. Evidence for the involvement of miRNA in redox regulated angiogenic response of human microvascular endothelial cells. Arteriosclerosis, thrombosis, and vascular biology. 2008;28(3):471-7.
123. Cheng Y, Liu X, Zhang S, Lin Y, Yang J, Zhang C. MicroRNA-21 protects against the H(2)O(2)-induced injury on cardiac myocytes via its target gene PDCD4. Journal of molecular and cellular cardiology. 2009;47(1):5-14.
124. Yousef M, Showe L, Showe M. A study of microRNAs in silico and in vivo: bioinformatics approaches to microRNA discovery and target identification. The FEBS journal. 2009;276(8):2150-6.
125. Luo R, Wong T, Zhu J, Liu CM, Zhu X, Wu E, Lee LK, Lin H, Zhu W, Cheung DW, Ting HF, Yiu SM, Peng S, Yu C, Li Y, Li R, Lam TW. SOAP3-dp: fast, accurate and sensitive GPU-based short read aligner. PloS one. 2013;8(5):e65632.
126. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research. 2010;20(9):1297-303.
127. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature protocols. 2009;4(7):1073-81.
128. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. Nature methods. 2010;7(4):248-9.
129. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Genome research. 2009;19(9):1553-61.
130. Schwarz JM, Rodelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. Nature methods. 2010;7(8):575-6.
131. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome research. 2010;20(1):110-21.
132. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, Lin J, Minguez P, Bork P, von Mering C, Jensen LJ. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic acids research. 2013;41(Database issue):D808-15.
133. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research. 2003;13(11):2498-504.
134. Mi H, Muruganujan A, Thomas PD. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucleic acids research. 2013;41(Database issue):D377-86.

135. Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, Gil L, Garcia-Giron C, Gordon L, Hourlier T, Hunt S, Juettemann T, Kahari AK, Keenan S, Komorowska M, Kulesha E, Longden I, Maurel T, McLaren WM, Muffato M, Nag R, Overduin B, Pignatelli M, Pritchard B, Pritchard E, Riat HS, Ritchie GR, Ruffier M, Schuster M, Sheppard D, Sobral D, Taylor K, Thormann A, Trevanion S, White S, Wilder SP, Aken BL, Birney E, Cunningham F, Dunham I, Harrow J, Herrero J, Hubbard TJ, Johnson N, Kinsella R, Parker A, Spudich G, Yates A, Zadissa A, Searle SM. Ensembl 2013. Nucleic acids research. 2013;41(Database issue):D48-55.
136. Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in bioinformatics. 2013;14(2):178-92.
137. Thorn CF, Klein TE, Altman RB. PharmGKB: the pharmacogenetics and pharmacogenomics knowledge base. Methods Mol Biol. 2005;311:179-91.
138. Guo L, Du Y, Chang S, Zhang K, Wang J. rSNPBase: a database for curated regulatory SNPs. Nucleic acids research. 2014;42(Database issue):D1033-9.
139. Kutmon M, Kelder T, Mandaviya P, Evelo CT, Coort SL. CyTargetLinker: a cytoscape app to integrate regulatory interactions in network analysis. PloS one. 2013;8(12):e82160.
140. Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, Bodmer J, Bodmer WF, Bonne-Tamir B, Cambon-Thomsen A, Chen Z, Chu J, Carcassi C, Contu L, Du R, Excoffier L, Ferrara GB, Friedlaender JS, Groot H, Gurwitz D, Jenkins T, Herrera RJ, Huang X, Kidd J, Kidd KK, Langaney A, Lin AA, Mehdi SQ, Parham P, Piazza A, Pistillo MP, Qian Y, Shu Q, Xu J, Zhu S, Weber JL, Greely HT, Feldman MW, Thomas G, Dausset J, Cavalli-Sforza LL. A human genome diversity cell line panel. Science. 2002;296(5566):261-2.
141. Illumina I. Quality Scores for Next-generation Sequencing. Assessing sequencing accuracy using Phred quality scoring. San Diego, CA USA: 2011.
142. Patowary A, Purkanti R, Singh M, Chauhan RK, Bhartiya D, Dwivedi OP, Chauhan G, Bharadwaj D, Sivasubbu S, Scaria V. Systematic analysis and functional annotation of variations in the genome of an Indian individual. Human mutation. 2012;33(7):1133-40.
143. Montgomery SB, Goode DL, Kvikstad E, Albers CA, Zhang ZD, Mu XJ, Ananda G, Howie B, Karczewski KJ, Smith KS, Anaya V, Richardson R, Davis J, MacArthur DG, Sidow A, Duret L, Gerstein M, Makova KD, Marchini J, McVean G, Lunter G. The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. Genome research. 2013;23(5):749-61.
144. Sjodin P, Bataillon T, Schierup MH. Insertion and deletion processes in recent human history. PloS one. 2010;5(1):e8650.
145. Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. A fine-scale map of recombination rates and hotspots across the human genome. Science. 2005;310(5746):321-4.
146. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, Zhou J, Gabriel SB, Barry R, Blumenstiel B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao Z, Huang W, Chu X, He Y, Jin L, Liu Y, Sun W, Wang H, Wang Y, Xiong X, Xu L, Waye MM, Tsui SK, Xue H, Wong JT, Galver LM, Fan JB, Gunderson K, Murray SS, Oliphant AR, Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier JF, Phillips MS, Roumy S, Sallee C, Verner A, Hudson TJ, Kwok PY, Cai

D, Koboldt DC, Miller RD, Pawlikowska L, Taillon-Miller P, Xiao M, Tsui LC, Mak W, Song YQ, Tam PK, Nakamura Y, Kawaguchi T, Kitamoto T, Morizono T, Nagashima A, Ohnishi Y, Sekine A, Tanaka T, Tsunoda T, Deloukas P, Bird CP, Delgado M, Dermitzakis ET, Gwilliam R, Hunt S, Morrison J, Powell D, Stranger BE, Whittaker P, Bentley DR, Daly MJ, de Bakker PI, Barrett J, Chretien YR, Maller J, McCarroll S, Patterson N, Pe'er I, Price A, Purcell S, Richter DJ, Sabeti P, Saxena R, Schaffner SF, Sham PC, Varilly P, Stein LD, Krishnan L, Smith AV, Tello-Ruiz MK, Thorisson GA, Chakravarti A, Chen PE, Cutler DJ, Kashuk CS, Lin S, Abecasis GR, Guan W, Li Y, Munro HM, Qin ZS, Thomas DJ, McVean G, Auton A, Bottolo L, Cardin N, Eyheramendy S, Freeman C, Marchini J, Myers S, Spencer C, Stephens M, Donnelly P, Cardon LR, Clarke G, Evans DM, Morris AP, Weir BS, Mullikin JC, Sherry ST, Feolo M, Skol A, Zhang H, Matsuda I, Fukushima Y, Macer DR, Suda E, Rotimi CN, Adebamowo CA, Ajayi I, Aniagwu T, Marshall PA, Nkwodimmah C, Royal CD, Leppert MF, Dixon M, Peiffer A, Qiu R, Kent A, Kato K, Niikawa N, Adewole IF, Knoppers BM, Foster MW, Clayton EW, Watkin J, Muzny D, Nazareth L, Sodergren E, Weinstock GM, Yakub I, Birren BW, Wilson RK, Fulton LL, Rogers J, Burton J, Carter NP, Clee CM, Griffiths M, Jones MC, McLay K, Plumb RW, Ross MT, Sims SK, Willey DL, Chen Z, Han H, Kang L, Godbout M, Wallenburg JC, L'Archeveque P, Bellemare G, Saeki K, An D, Fu H, Li Q, Wang Z, Wang R, Holden AL, Brooks LD, McEwen JE, Guyer MS, Wang VO, Peterson JL, Shi M, Spiegel J, Sung LM, Zacharia LF, Collins FS, Kennedy K, Jamieson R, Stewart J. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007;449(7164):851-61.

147. Nachman MW. Single nucleotide polymorphisms and recombination rate in humans. Trends in genetics : TIG. 2001;17(9):481-5.
148. Zhao Z, Fu YX, Hewett-Emmett D, Boerwinkle E. Investigating single nucleotide polymorphism (SNP) density in the human genome and its implications for molecular evolution. Gene. 2003;312:207-13.
149. Nielsen R. Molecular signatures of natural selection. Annual review of genetics. 2005;39:197-218.
150. Gorlov IP, Kimmel M, Amos CI. Strength of the purifying selection against different categories of the point mutations in the coding regions of the human genome. Human molecular genetics. 2006;15(7):1143-50.
151. Tong P, Prendergast JG, Lohan AJ, Farrington SM, Cronin S, Friel N, Bradley DG, Hardiman O, Evans A, Wilson JF, Loftus B. Sequencing and analysis of an Irish human genome. Genome biology. 2010;11(9):R91.
152. Petrov DA. Mutational equilibrium model of genome size evolution. Theoretical population biology. 2002;61(4):531-44.
153. Taylor MS, Kai C, Kawai J, Carninci P, Hayashizaki Y, Semple CA. Heterotachy in mammalian promoter evolution. PLoS genetics. 2006;2(4):e30.
154. Huang S, Li J, Xu A, Huang G, You L. Small insertions are more deleterious than small deletions in human genomes. Human mutation. 2013;34(12):1642-9.
155. Zhao H, Li Q, Li J, Zeng C, Hu S, Yu J. The study of neighboring nucleotide composition and transition/transversion bias. Science in China Series C, Life sciences / Chinese Academy of Sciences. 2006;49(4):395-402.
156. Watson JD. Molecular biology of the gene. 6th ed. San Francisco Cold Spring Harbor, N.Y.: Pearson/Benjamin Cummings; Cold Spring Harbor Laboratory Press; 2008. xxxii, 841 p. p.
157. Ruvinsky A. Genetics and randomness. Boca Raton, FL: CRC Press; 2010. xiii, 169 p. p.

158. Shen H, Li J, Zhang J, Xu C, Jiang Y, Wu Z, Zhao F, Liao L, Chen J, Lin Y, Tian Q, Papasian CJ, Deng HW. Comprehensive characterization of human genome variation by high coverage whole-genome sequencing of forty four Caucasians. PloS one. 2013;8(4):e59494.
159. Jobling MA, Tyler-Smith C. The human Y chromosome: an evolutionary marker comes of age. Nature reviews Genetics. 2003;4(8):598-612.
160. Wilson Sayres MA, Lohmueller KE, Nielsen R. Natural selection reduced diversity on human y chromosomes. PLoS genetics. 2014;10(1):e1004064.
161. Liu X, Jian X, Boerwinkle E. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. Human mutation. 2011;32(8):894-9.
162. Calin GA, Liu CG, Ferracin M, Hyslop T, Spizzo R, Sevignani C, Fabbri M, Cimmino A, Lee EJ, Wojcik SE, Shimizu M, Tili E, Rossi S, Taccioli C, Pichiorri F, Liu X, Zupo S, Herlea V, Gramantieri L, Lanza G, Alder H, Rassenti L, Volinia S, Schmittgen TD, Kipps TJ, Negrini M, Croce CM. Ultraconserved regions encoding ncRNAs are altered in human leukemias and carcinomas. Cancer cell. 2007;12(3):215-29.
163. Devlin TM. Textbook of biochemistry : with clinical correlations. 7th ed. Hoboken, NJ: John Wiley & Sons; 2011. xxxii, 1204 p. p.
164. Mochizuki T, Lemmink HH, Mariyama M, Antignac C, Gubler MC, Pirson Y, Verellen-Dumoulin C, Chan B, Schroder CH, Smeets HJ, et al. Identification of mutations in the alpha 3(IV) and alpha 4(IV) collagen genes in autosomal recessive Alport syndrome. Nature genetics. 1994;8(1):77-81.
165. Gupta S, Batchu RB, Datta K. Purification, partial characterization of rat kidney hyaluronic acid binding protein and its localization on the cell surface. European journal of cell biology. 1991;56(1):58-67.
166. Willeit J, Kiechl S, Weimer T, Mair A, Santer P, Wiedermann CJ, Roemisch J. Marburg I polymorphism of factor VII--activating protease: a prominent risk predictor of carotid stenosis. Circulation. 2003;107(5):667-70.
167. Hoppe B, Tolou F, Radtke H, Kiesewetter H, Dorner T, Salama A. Marburg I polymorphism of factor VII-activating protease is associated with idiopathic venous thromboembolism. Blood. 2005;105(4):1549-51.
168. Graf J, Voisey J, Hughes I, van Daal A. Promoter polymorphisms in the MATP (SLC45A2) gene are associated with normal human skin color variation. Human mutation. 2007;28(7):710-7.
169. Yuasa I, Umetsu K, Harihara S, Kido A, Miyoshi A, Saitou N, Dashnyam B, Jin F, Lucotte G, Chattopadhyay PK, Henke L, Henke J. Distribution of the F374 allele of the SLC45A2 (MATP) gene and founder-haplotype analysis. Annals of human genetics. 2006;70(Pt 6):802-11.
170. Guedj M, Bourillon A, Combadieres C, Rodero M, Dieude P, Descamps V, Dupin N, Wolkenstein P, Aegerter P, Lebbe C, Basset-Seguin N, Prum B, Saiag P, Grandchamp B, Soufir N. Variants of the MATP/SLC45A2 gene are protective for melanoma in the French population. Human mutation. 2008;29(9):1154-60.
171. Fernandez LP, Milne RL, Pita G, Aviles JA, Lazaro P, Benitez J, Ribas G. SLC45A2: a novel malignant melanoma-associated gene. Human mutation. 2008;29(9):1161-7.
172. Stacey SN, Sulem P, Masson G, Gudjonsson SA, Thorleifsson G, Jakobsdottir M, Sigurdsson A, Gudbjartsson DF, Sigurgeirsson B, Benediktsdottir KR, Thorisdottir K, Ragnarsson R, Scherer D, Hemminki K, Rudnai P, Gurzau E, Koppova K, Botella-Estrada R, Soriano V, Juberias P, Saez B, Gilaberte Y, Fuentelsaz V, Corredera C, Grasa M, Hoiom V, Lindblom A,

Bonenkamp JJ, van Rossum MM, Aben KK, de Vries E, Santinami M, Di Mauro MG, Maurichi A, Wendt J, Hochleitner P, Pehamberger H, Gudmundsson J, Magnusdottir DN, Gretarsdottir S, Holm H, Steinthorsdottir V, Frigge ML, Blondal T, Saemundsdottir J, Bjarnason H, Kristjansson K, Bjornsdottir G, Okamoto I, Rivoltini L, Rodolfo M, Kiemeney LA, Hansson J, Nagore E, Mayordomo JI, Kumar R, Karagas MR, Nelson HH, Gulcher JR, Rafnar T, Thorsteinsdottir U, Olafsson JH, Kong A, Stefansson K. New common variants affecting susceptibility to basal cell carcinoma. Nature genetics. 2009;41(8):909-14.

173. Woo HI, Kim KK, Choi H, Kim S, Jang KT, Yi JH, Park YS, Park JO, Lee SY. Effect of genetic polymorphisms on therapeutic response and clinical outcomes in pancreatic cancer patients treated with gemcitabine. Pharmacogenomics. 2012;13(9):1023-35.
174. Skipper L, Li Y, Bonnard C, Pavanni R, Yih Y, Chua E, Sung WK, Tan L, Wong MC, Tan EK, Liu J. Comprehensive evaluation of common genetic variation within LRRK2 reveals evidence for association with sporadic Parkinson's disease. Human molecular genetics. 2005;14(23):3549-56.
175. Tian JY, Guo JF, Wang L, Sun QY, Yao LY, Luo LZ, Shi CH, Hu YC, Yan XX, Tang BS. Mutation analysis of LRRK2, SCNA, UCHL1, HtrA2 and GIGYF2 genes in Chinese patients with autosomal dorminant Parkinson's disease. Neuroscience letters. 2012;516(2):207-11.
176. Martignoni M, Riano E, Rugarli EI. The role of ZFYVE27/protrudin in hereditary spastic paraplegia. American journal of human genetics. 2008;83(1):127-8; author reply 8-30.
177. McGuire MC, Nogueira CP, Bartels CF, Lightstone H, Hajra A, Van der Spek AF, Lockridge O, La Du BN. Identification of the structural mutation responsible for the dibucaine-resistant (atypical) variant form of human serum cholinesterase. Proceedings of the National Academy of Sciences of the United States of America. 1989;86(3):953-7.
178. Clementi R, Emmi L, Maccario R, Liotta F, Moretta L, Danesino C, Arico M. Adult onset and atypical presentation of hemophagocytic lymphohistiocytosis in siblings carrying PRF1 mutations. Blood. 2002;100(6):2266-7.
179. Perez de Diego R, Sancho-Shimizu V, Lorenzo L, Puel A, Plancoulaine S, Picard C, Herman M, Cardon A, Durandy A, Bustamante J, Vallabhapurapu S, Bravo J, Warnatz K, Chaix Y, Cascarrigny F, Lebon P, Rozenberg F, Karin M, Tardieu M, Al-Muhsen S, Jouanguy E, Zhang SY, Abel L, Casanova JL. Human TRAF3 adaptor molecule deficiency leads to impaired Toll-like receptor 3 response and susceptibility to herpes simplex encephalitis. Immunity. 2010;33(3):400-11.
180. Khrunin A, Ivanova F, Moisseev A, Khokhrin D, Sleptsova Y, Gorbunova V, Limborska S. Pharmacogenomics of cisplatin-based chemotherapy in ovarian cancer patients of different ethnic origins. Pharmacogenomics. 2012;13(2):171-8.
181. Williams EL, Acquaviva C, Amoroso A, Chevalier F, Coulter-Mackie M, Monico CG, Giachino D, Owen T, Robbiano A, Salido E, Waterham H, Rumsby G. Primary hyperoxaluria type 1: update and additional mutation analysis of the AGXT gene. Human mutation. 2009;30(6):910-7.
182. Cecchin E, D'Andrea M, Lonardi S, Zanusso C, Pella N, Errante D, De Mattia E, Polesel J, Innocenti F, Toffoli G. A prospective validation pharmacogenomic study in the adjuvant setting of colorectal cancer patients treated with the 5-fluorouracil/leucovorin/oxaliplatin (FOLFOX4) regimen. The pharmacogenomics journal. 2013;13(5):403-9.
183. Goswami S, Gong L, Giacomini K, Altman RB, Klein TE. PharmGKB summary: very important pharmacogene information for SLC22A1. Pharmacogenetics and genomics. 2014;24(6):324-8.

184. Zhang J, Chalmers MJ, Stayrook KR, Burris LL, Wang Y, Busby SA, Pascal BD, Garcia-Ordonez RD, Bruning JB, Istrate MA, Kojetin DJ, Dodge JA, Burris TP, Griffin PR. DNA binding alters coactivator interaction surfaces of the intact VDR-RXR complex. Nature structural & molecular biology. 2011;18(5):556-63.
185. Baudino TA, Kraichely DM, Jefcoat SC, Jr., Winchester SK, Partridge NC, MacDonald PN. Isolation and characterization of a novel coactivator protein, NCoA-62, involved in vitamin D-mediated transcription. The Journal of biological chemistry. 1998;273(26):16434-41.
186. Chen WY, Bertone-Johnson ER, Hunter DJ, Willett WC, Hankinson SE. Associations between polymorphisms in the vitamin D receptor and breast cancer risk. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2005;14(10):2335-9.
187. Creatsa M, Pliatsika P, Kaparos G, Antoniou A, Armeni E, Tsakonas E, Panoulis C, Alexandrou A, Dimitraki E, Christodoulakos G, Lambrinoudaki I. The effect of vitamin D receptor BsmI genotype on the response to osteoporosis treatment in postmenopausal women: a pilot study. The journal of obstetrics and gynaecology research. 2011;37(10):1415-22.
188. Morrison NA, George PM, Vaughan T, Tilyard MW, Frampton CM, Gilchrist NL. Vitamin D receptor genotypes influence the success of calcitriol therapy for recurrent vertebral fracture in osteoporosis. Pharmacogenetics and genomics. 2005;15(2):127-35.
189. Oakley-Girvan I, Feldman D, Eccleshall TR, Gallagher RP, Wu AH, Kolonel LN, Halpern J, Balise RR, West DW, Paffenbarger RS, Jr., Whittemore AS. Risk of early-onset prostate cancer in relation to germ line polymorphisms of the vitamin D receptor. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2004;13(8):1325-30.
190. Roth DE, Soto G, Arenas F, Bautista CT, Ortiz J, Rodriguez R, Cabrera L, Gilman RH. Association between vitamin D receptor gene polymorphisms and response to treatment of pulmonary tuberculosis. The Journal of infectious diseases. 2004;190(5):920-7.
191. WHO. STOPTB Partnership – Tuberculosis Global Facts 2011/2012. 2011.
192. Saúde MdSP-DGd. Programa Nacional de Luta Contra a Tuberculose: Relatório para o Dia Mundial da Tuberculose 2013. Ponto da Situação Epidemiológica de Desempenho. Direção Geral da Saúde, 2013.
193. WHO. World Health Organization – TB/HIV Manual Clínico 2nd Ed. ed. WHO, editor. Geneva2005.
194. Wilkinson RJ, Llewelyn M, Toossi Z, Patel P, Pasvol G, Lalvani A, Wright D, Latif M, Davidson RN. Influence of vitamin D deficiency and vitamin D receptor polymorphisms on tuberculosis among Gujarati Asians in west London: a case-control study. Lancet. 2000;355(9204):618-21.
195. Liu PT, Stenger S, Li H, Wenzel L, Tan BH, Krutzik SR, Ochoa MT, Schauber J, Wu K, Meinken C, Kamen DL, Wagner M, Bals R, Steinmeyer A, Zugel U, Gallo RL, Eisenberg D, Hewison M, Hollis BW, Adams JS, Bloom BR, Modlin RL. Toll-like receptor triggering of a vitamin D-mediated human antimicrobial response. Science. 2006;311(5768):1770-3.
196. Azad AK, Sadee W, Schlesinger LS. Innate immune gene polymorphisms in tuberculosis. Infection and immunity. 2012;80(10):3343-59.

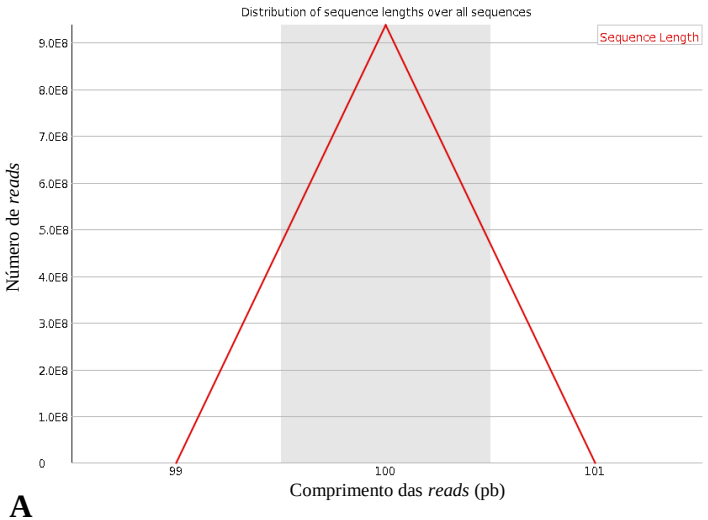
197. Chen R, Davydov EV, Sirota M, Butte AJ. Non-synonymous and synonymous coding SNPs show similar likelihood and effect size of human disease association. PloS one. 2010;5(10):e13574.
198. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences of the United States of America. 2009;106(23):9362-7.
199. Glisovic T, Bachorik JL, Yong J, Dreyfuss G. RNA-binding proteins and post-transcriptional gene regulation. FEBS letters. 2008;582(14):1977-86.
200. Lahr MM, Foley RA. Towards a theory of modern human origins: geography, demography, and diversity in recent human evolution. American journal of physical anthropology. 1998;Suppl 27:137-76.
201. Underhill PA, Passarino G, Lin AA, Shen P, Mirazon Lahr M, Foley RA, Oefner PJ, Cavalli-Sforza LL. The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. Annals of human genetics. 2001;65(Pt 1):43-62.
202. Templeton A. Out of Africa again and again. Nature. 2002;416(6876):45-51.
203. Stringer C. Palaeoanthropology. Coasting out of Africa. Nature. 2000;405(6782):24-5, 7.
204. Fagan BM. The great journey : the peopling of ancient America. Updated ed. Gainesville: University Press of Florida; 2004. xxiii, 288 p. p.
205. Cavalli-Sforza LL. Genes, peoples, and languages. 1st ed. New York: North Point Press; 2000. xii, 227 p. p.

Anexos

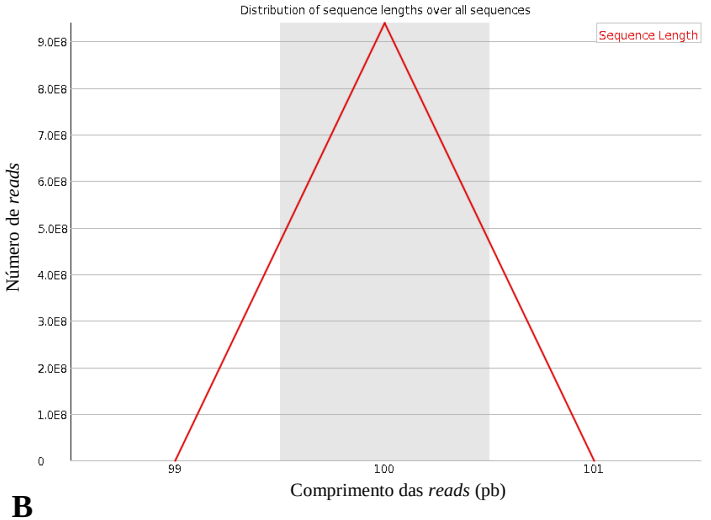
Anexo 1: Descrição das regiões genómicas onde foram anotadas as variações pontuais, nos genomas H1, H2, H3 e H4.

5'UTR		Variantes que se localizam na região não traduzida 5'.
3'UTR		Variantes que se localizam na região não traduzida 3'.
5'UTR_3'UTR		Variantes que se localizam em ambas as regiões 5'UTR e 3'UTR (possivelmente para dois genes diferentes).
Upstream		Variantes que se localizam numa região de 1kb a montante do local de iniciação da transcrição.
Downstream		Variantes que se localizam numa região de 1kb a jusante do local final da transcrição.
Upstream_downstream		Variantes que se localizam em ambas as regiões *upstream* e *downstream* (possivelmente para dois genes diferentes).
Intrónica		Variantes que se localizam numa região de intrão.
Exónica		Variantes que se localizam numa região de exão codificante.
Splicing		Variantes que se localizam dentro dos 2 pb de distância a partir de uma fronteira exão/intrão.
Exónica_*Splicing*		Variantes que se localizam dentro de um exão mas na proximidade de uma fronteira exão/intrão.
ncRNA_	**5'UTR**	Variantes que se localizam nas regiões: 5'UTR, 3'UTR, *splicing*, intrónica e exónica; referentes a genes cujo RNA não contém anotação codificante conhecida na versão do ANNOVAR utilizada neste estudo.
	3'UTR	
	Splicing	
	Intrónica	
	Exónica	
Intergénica		Variantes que se localizam numa região intergénica.

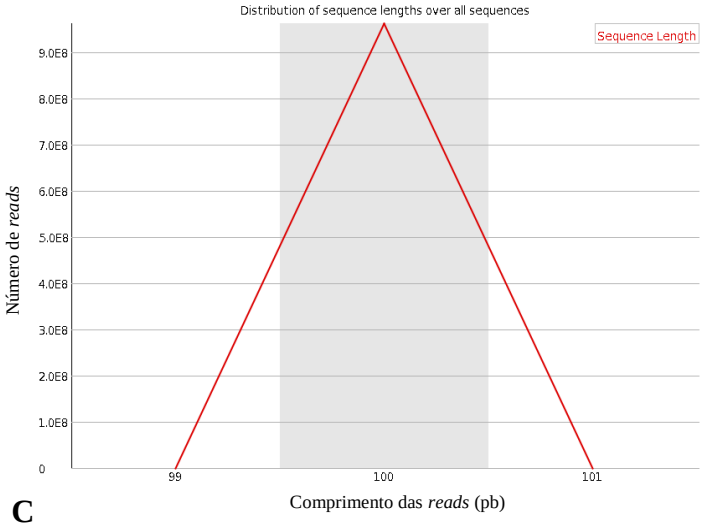
Anexo 2: Comprimento das *reads* dos genomas H1 (A), H2 (B), H3 (C) e H4 (D), analisado pelo FastQC.



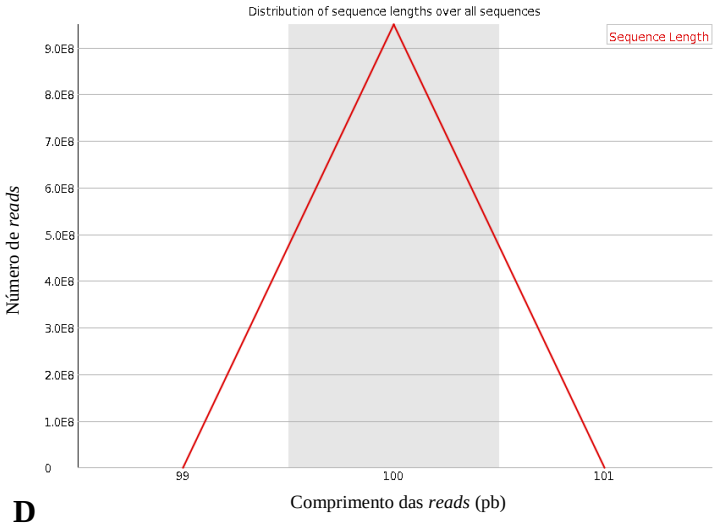
A



B

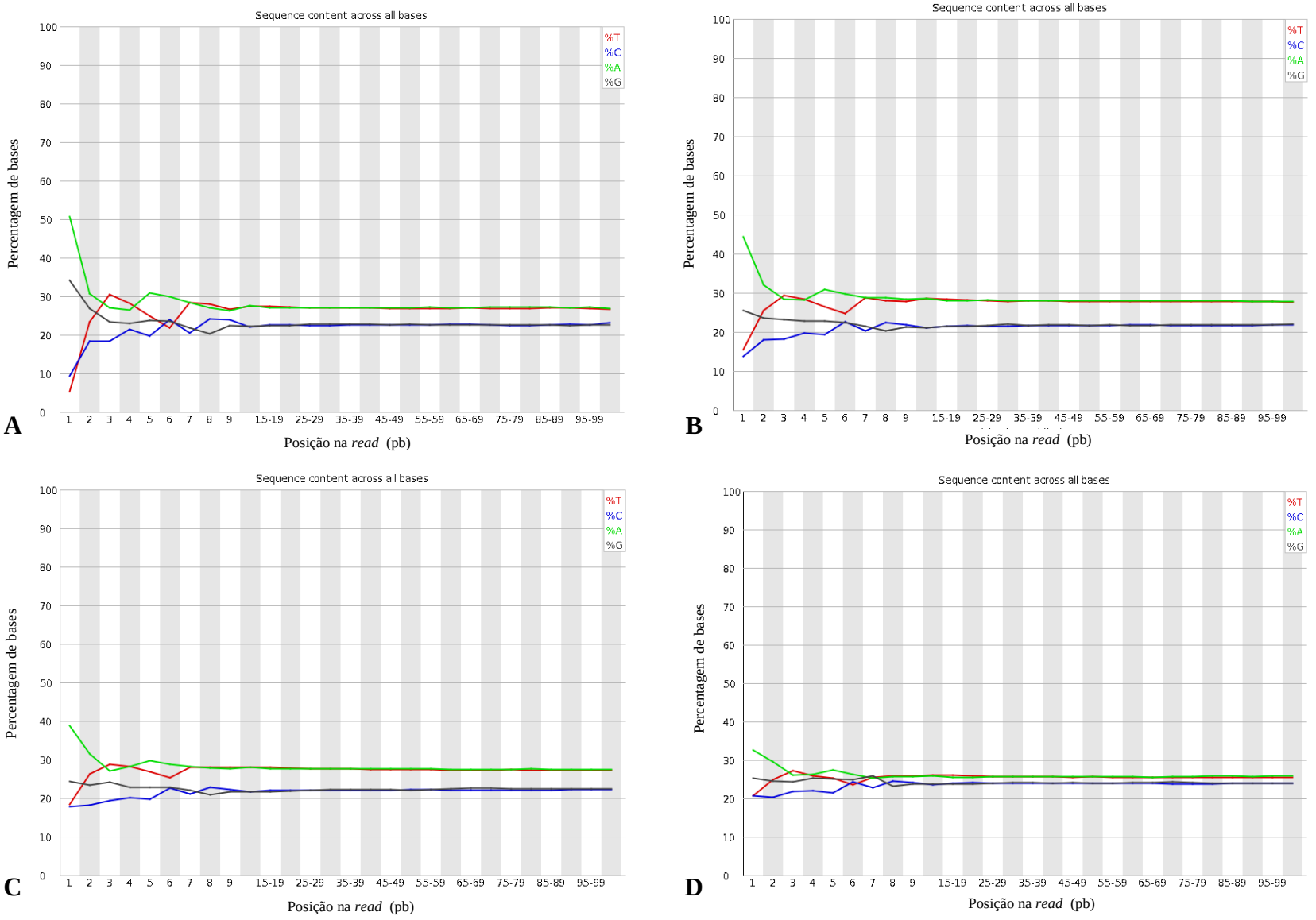


C

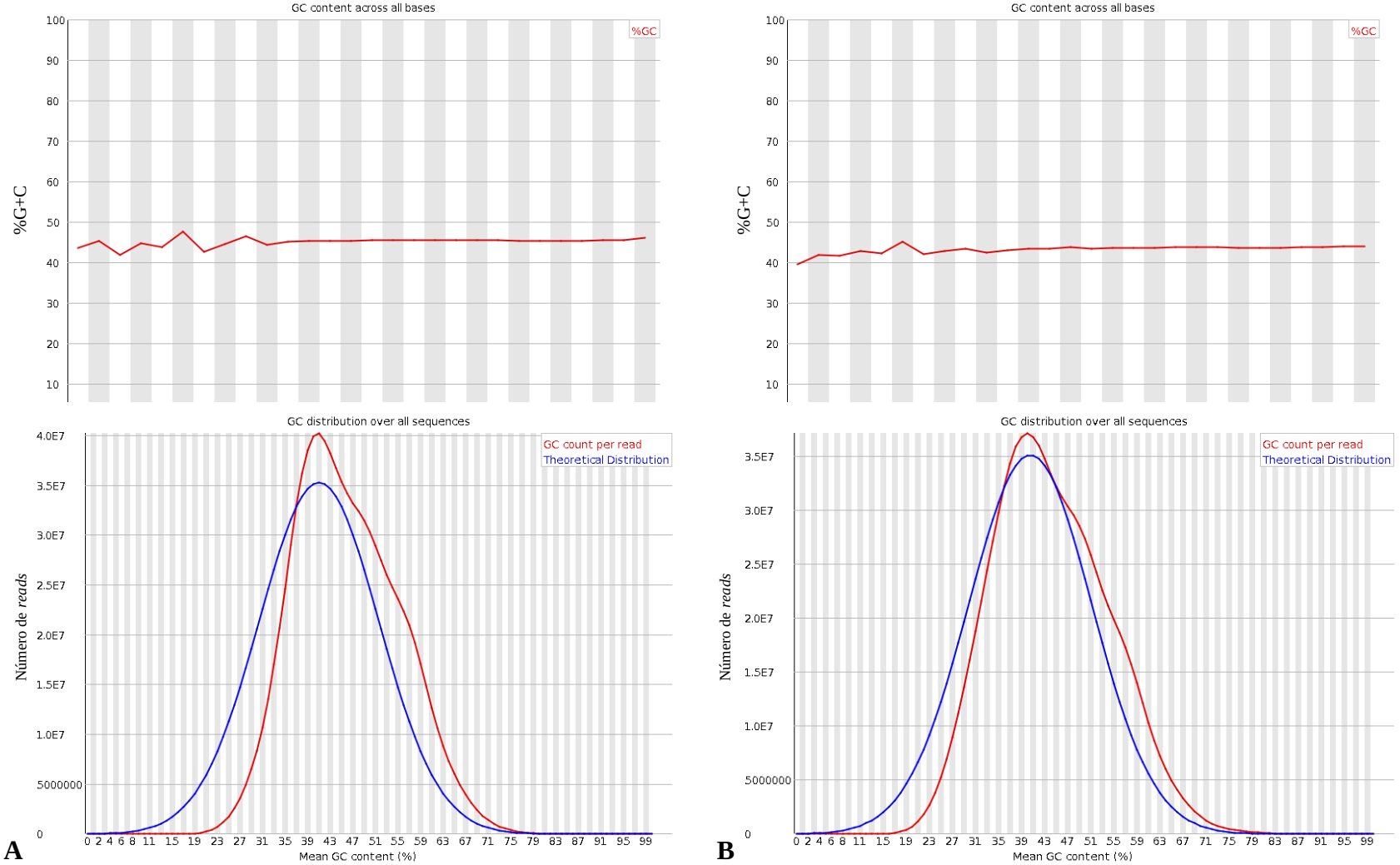


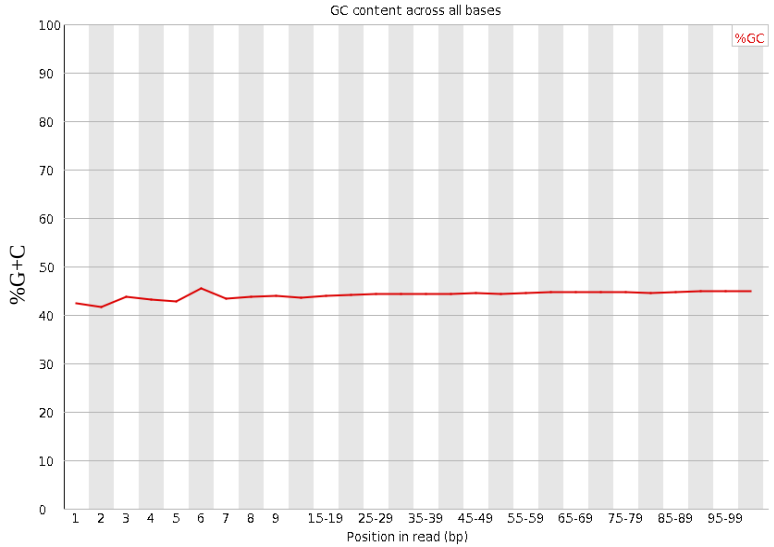
D

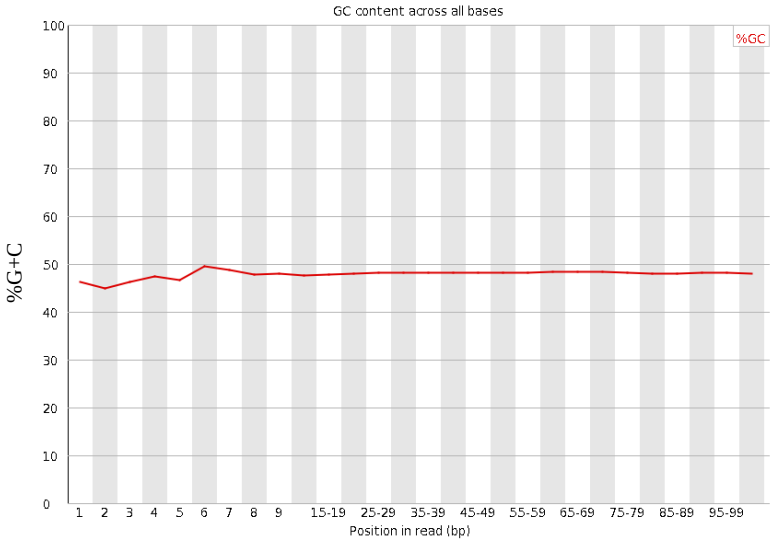
Anexo 3: Proporção média das bases sequenciadas em cada posição das *reads* dos genomas H1, H2, H3 e H4, analisado pelo FastQC.

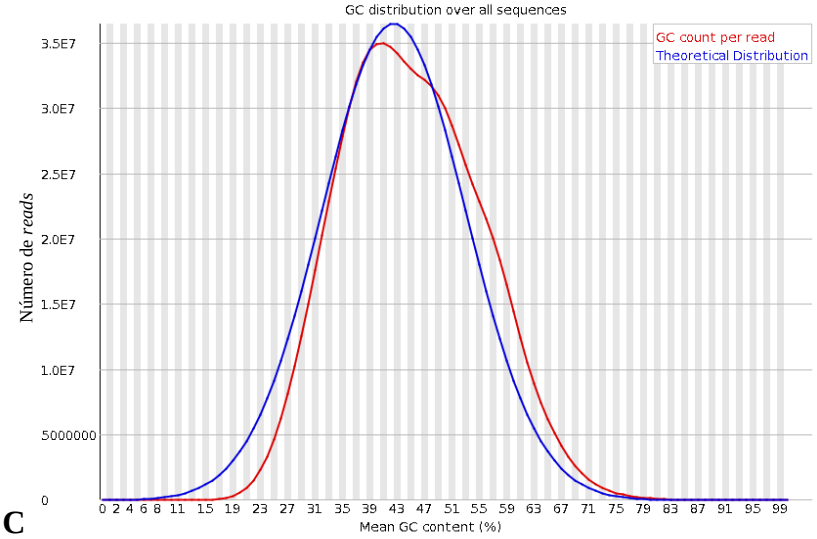


Anexo 4: Conteúdo em %G+C nas *reads* dos genomas H1 (A), H2 (B), H3 (C) e H4 (D), analisado pelo FastQC.

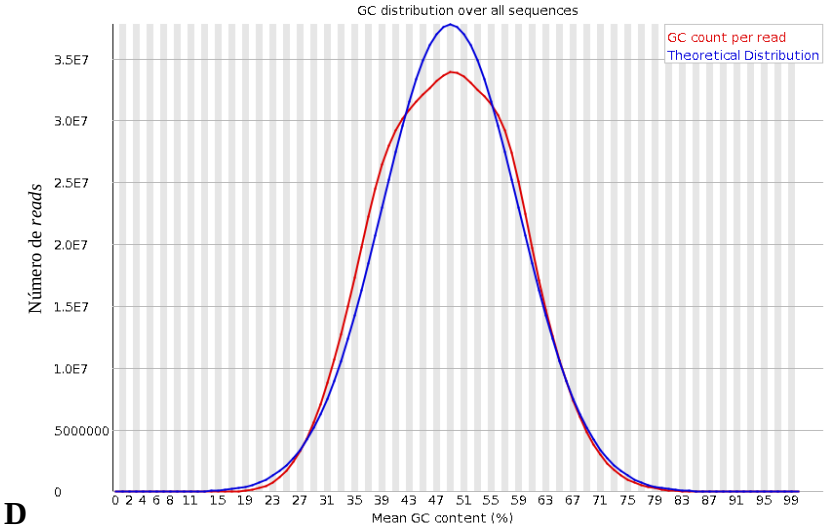



GC content across all bases
%GC
%G+C
Position in read (bp)


GC content across all bases
%GC
%G+C
Position in read (bp)


GC distribution over all sequences
GC count per read
Theoretical Distribution
Número de reads
Mean GC content (%)

C


GC distribution over all sequences
GC count per read
Theoretical Distribution
Número de reads
Mean GC content (%)

D

Anexo 5: Variantes filtradas pelo SIFT, PolyPhen 2, LRT e *Mutation Taster*, nos genomas H1, H2, H3 e H4.

H1 Cromossoma	Localização	Ref	Alt	Gene	ESP (americanos europeus)	1000 Genomas (população mundial)	dbSNP137
1	32797330	T	A	*HDAC1*	NA	NA	NA
1	47834209	G	T	*CMPK1*	0.048256	0.06	rs35687416
1	151337045	C	T	*SELENBP1*	NA	0.0005	rs116396590
1	154942911	G	A	*SHC1*	0.002020	0.0018	rs115641580
1	183849869	C	G	*RGL1*	0.003721	0.0018	rs12140352
1	201017805	A	C	*CACNA1S*	NA	NA	rs79011683
1	206224635	G	C	*AVPR1B*	0.061279	0.02	rs35369693
1	208272313	A	C	*PLXNA2*	NA	NA	rs200698765
1	203186947	G	C	*CHIT1*	NA	NA	rs201682373
1	36225948	T	C	*CLSPN*	0.133488	0.46	rs7537203
1	36202585	G	A	*CLSPN*	0.064884	0.05	rs35490896
1	229568310	C	A	*ACTA1*	NA	NA	NA
1	235894366	A	C	*LYST*	0.003721	0.0009	rs34702903
2	3392295	A	G	*TTC15*	0.374070	0.44	rs11686212
2	18112542	T	A	*KCNS3*	NA	NA	NA
2	26712094	C	T	*OTOF*	NA	NA	rs200958458
2	74759825	G	A	*HTRA2*	0.004884	0.0023	NA
2	105859249	C	T	*GPR45*	0.145698	0.10	rs35946826
2	150432277	A	G	*MMADHC*	0.000467	0	rs61755260
2	217311766	C	T	*SMARCAL1*	NA	NA	NA
2	219256102	G	A	*SLC11A1*	0.002675	0.0014	rs141861983
2	228163475	G	A	*COL4A3*	0.000243	0.0014	rs190598500
2	230861519	G	T	*FBXO36*	0.240000	0.21	rs1035834
2	233346498	C	T	*ECEL1*	0.001860	0.01	rs142492002
3	14724345	G	A	*C3orf20*	0.058605	0.06	rs17040154
3	38167095	A	G	*ACAA1*	0.040698	0.02	rs2229528
3	58183636	G	A	*DNASE1L3*	0.075116	0.03	rs35677470
3	67053926	T	C	*KBTBD8*	0.066047	0.03	rs75804175
3	99885236	A	G	*C3orf26*	0.022214	0.01	rs11537817
3	118942956	C	A	*B4GALT4*	0.000349	NA	rs138943942
3	196771513	G	A	*DLG1*	0.047093	0.03	rs34492126
4	95220779	A	G	*HPGDS*	0.000814	0.0009	rs146652732
4	102751014	G	C	*BANK1*	0.017907	0.01	rs35978636
4	106859549	G	C	*NPNT*	0.325233	0.32	rs35132891
5	32255800	G	C	*MTMR12*	NA	NA	NA
5	33951693	C	G	*SLC45A2*	0.959186	0.44	rs16891982
5	86695274	A	G	*CCNH*	0.223837	0.14	rs2230641
5	141019569	G	C	*RELL2*	0.194580	0.16	rs17855844

H1 Cromossoma	Localização	Ref	Alt	Gene	ESP (americanos europeus)	1000 Genomas (população mundial)	dbSNP137
5	141019830	C	G	*RELL2*	0.196860	0.16	rs11742646
5	149301223	G	C	*PDE6A*	NA	0.01	rs61733363
5	96232560	T	A	*ERAP2*	NA	NA	NA
5	118469561	G	A	*DMXL1*	0.001279	NA	rs139365266
5	120021817	C	A	*PRR16*	0.158488	0.10	rs17853861
6	31737454	G	A	*C6orf27*	NA	NA	NA
6	4087949	A	T	*C6orf201*	0.066304	0.04	rs13200786
6	112508770	G	T	*LAMA4*	NA	0.82	rs9400522
6	112522852	G	A	*LAMA4*	0.064884	0.03	rs11757455
6	116783330	G	A	*FAM26F*	0.186991	0.23	rs1057192
6	136560616	C	T	*FAM54A*	0.002209	0.0014	rs145047825
6	76022171	G	A	*FILIP1*	0.002442	0.01	rs35227190
6	116943975	G	A	*RSPH4A*	0.023256	0.01	rs41289942
6	31778263	C	A	*HSPA1L*	NA	NA	NA
6	97562116	C	T	*KLHL32*	NA	NA	NA
6	53519740	C	T	*KLHL31*	0.001279	0.0005	rs141075656
6	117113315	G	A	*GPRC6A*	0.024070	0.01	rs41290852
7	130002797	G	A	*CPA5*	NA	0.0023	rs74947917
7	117171037	G	A	*CFTR*	NA	NA	rs201958172
7	87913221	C	T	*STEAP4*	0.199637	0.10	rs34741656
7	103008231	A	T	*PSMC2*	NA	NA	NA
7	56087374	C	T	*PSPH*	NA	NA	rs200442078
8	56986648	G	T	*RPS20*	NA	NA	rs11537559
8	101718965	G	A	*PABPC1*	NA	NA	rs62513921
8	101724606	G	A	*PABPC1*	NA	NA	rs202060459
8	101724924	G	A	*PABPC1*	NA	NA	NA
8	101730043	G	C	*PABPC1*	NA	NA	rs113614781
8	101730073	T	C	*PABPC1*	NA	0.26	rs72681443
8	101730413	C	G	*PABPC1*	NA	NA	NA
8	101730432	A	T	*PABPC1*	NA	NA	NA
8	101730437	G	A	*PABPC1*	NA	NA	NA
9	368128	C	T	*DOCK8*	0.106047	0.04	rs17673268
9	1056870	C	T	*DMRT2*	0.015930	0.01	rs41311430
9	95007336	C	G	*IARS*	NA	NA	NA
9	95482883	T	C	*BICD2*	NA	NA	NA
9	116082647	C	G	*WDR31*	0.208953	0.10	rs41307479
9	125316028	T	C	*OR1N2*	0.096860	0.06	rs41297203
9	125486717	G	A	*OR1L4*	NA	NA	rs76170289
9	125512575	G	A	*OR1L6*	NA	0.92	rs4838012
10	12162877	C	A	*DHTKD1*	NA	NA	NA

H1 Cromossoma	Localização	Ref	Alt	Gene	ESP (americanos europeus)	1000 Genomas (população mundial)	dbSNP137
10	99509251	G	T	*ZFYVE27*	0.007674	0.02	rs35077384
10	102054823	G	A	*PKD2L1*	0.015698	0.01	rs147426900
10	105183348	T	C	*PDCD11*	0.013953	0.01	rs61751511
10	61802477	C	T	*ANK3*	0.006977	0.0032	rs141939315
10	113921450	C	A	*GPAM*	NA	NA	rs199856746
10	113935379	T	C	*GPAM*	0.587674	0.55	rs10787428
10	47087501	C	T	*PPYR1*	0.284884	0.30	rs3824733
10	47087609	G	A	*PPYR1*	0.013721	NA	rs79871698
10	115348046	G	A	*HABP2*	0.038837	0.01	rs7080536
10	126691575	T	C	*CTBP2*	NA	NA	rs112239066
10	126691579	C	T	*CTBP2*	NA	NA	NA
10	126691979	C	G	*CTBP2*	NA	NA	rs3198926
10	126692001	T	A	*CTBP2*	NA	NA	NA
10	126692013	A	G	*CTBP2*	NA	NA	NA
10	126692017	G	C	*CTBP2*	NA	NA	rs3198920
10	126692020	G	A	*CTBP2*	NA	NA	rs11550782
10	126692029	G	T	*CTBP2*	NA	NA	NA
10	126692037	C	A	*CTBP2*	NA	NA	NA
11	800449	G	A	*PIDD*	NA	NA	rs200290640
11	14496079	G	A	*COPB1*	0.000233	NA	rs142993682
11	55703011	G	A	*OR5I1*	0.053097	0.03	rs61995929
11	61252829	C	T	*C11orf66*	0.000233	NA	rs141041531
11	118985055	G	A	*C2CD2L*	NA	NA	NA
12	40707861	C	T	*LRRK2*	0.032915	0.02	rs33958906
12	48272895	A	G	*VDR*	0.614070	0.65	rs2228570
12	14664250	A	G	*PLBD1*	0.071163	0.06	rs2287541
12	50848132	T	A	*LARP4*	0.051326	0.03	rs17124706
12	52822136	C	G	*KRT75*	NA	NA	NA
12	58217422	T	C	*CTDSP2*	NA	NA	NA
12	58217453	G	A	*CTDSP2*	NA	NA	NA
12	58217698	G	T	*CTDSP2*	NA	NA	NA
12	70091452	A	G	*BEST3*	0.088156	0.04	rs1025016
12	122845698	C	A	*CLIP1*	NA	NA	rs79909185
12	124241506	C	T	*ATP6V0A2*	0.035233	0.02	rs17883456
13	24411715	G	A	*MIPEP*	0.000116	0.0005	rs188162736
13	114088074	G	A	*ADPRHL1*	0.024796	0.02	rs139075628
14	20876282	G	A	*TEP1*	0.021279	0.01	rs41310936
14	64596823	C	A	*SYNE2*	0.019767	0.0046	rs75568433
14	81380743	G	A	*CEP128*	NA	NA	NA
14	59950676	A	C	*C14orf149*	0.060970	0.04	rs34741399

H1 Cromossoma	Localização	Ref	Alt	Gene	ESP (americanos europeus)	1000 Genomas (população mundial)	dbSNP137
14	62229285	T	G	*SNAPC1*	0.006047	0.03	rs74810099
14	88946070	A	G	*PTPN21*	NA	NA	NA
15	43044464	C	A	*TTBK2*	NA	NA	NA
15	75500713	G	T	*C15orf39*	0.006636	0.01	rs79245819
15	91025227	T	G	*IQGAP1*	NA	NA	NA
15	91326099	C	T	*BLM*	0.065263	0.05	rs11852361
15	98512431	C	T	*ARRDC4*	0.043741	0.02	rs61747226
16	2027203	C	T	*TBL3*	NA	NA	NA
16	48151221	C	A	*ABCC12*	NA	NA	NA
16	81249927	C	T	*PKD1L2*	0.345524	0.18	rs7185774
16	81249954	T	A	*PKD1L2*	0.633445	0.52	rs7191351
16	67970188	G	C	*PSMB10*	NA	NA	NA
16	70563087	C	G	*SF3B3*	NA	NA	rs76371422
16	71683718	A	G	*PHLPP2*	0.149884	0.10	rs61733127
16	89704064	T	C	*DPEP1*	NA	NA	NA
18	57103297	T	G	*CCBE1*	0.002558	0.0005	rs139059968
17	2227655	G	T	*TSR1*	0.040000	0.01	rs35343613
17	17720319	C	T	*SREBF1*	0.015465	0.01	rs36215896
17	11672607	G	T	*DNAH9*	0.147442	0.08	rs61744697
17	18161945	G	T	*FLII*	0.022673	0.01	rs145840264
17	21204187	G	T	*MAP2K3*	NA	0.50	rs56067280
17	21204192	C	T	*MAP2K3*	NA	0.50	rs56216806
17	21217513	G	A	*MAP2K3*	NA	0.49	rs2363198
17	21319069	G	A	*KCNJ12*	NA	NA	rs76265595
17	21319079	C	A	*KCNJ12*	NA	NA	rs76518282
17	21319087	G	A	*KCNJ12*	NA	NA	rs75029097
17	21319121	C	T	*KCNJ12*	NA	0.50	rs1714864
17	41901366	G	T	*MPP3*	0.007383	0.0009	rs189143886
17	42927723	G	A	*HIGD1B*	0.023837	0.01	rs2231650
17	45216172	A	G	*CDC27*	NA	NA	rs76836152
17	45232043	T	C	*CDC27*	NA	NA	rs186452221
17	45235598	G	C	*CDC27*	NA	NA	rs77467652
17	45235635	A	C	*CDC27*	NA	NA	rs79936417
17	45249391	T	G	*CDC27*	NA	NA	rs62077270
17	48761053	G	A	*ABCC3*	0.063023	0.03	rs11568591
17	73238508	T	C	*GGA3*	0.009651	0.0018	rs52809447
19	8436373	C	T	*ANGPTL4*	0.003025	0.0009	rs140744493
19	39908214	C	T	*PLEKHG2*	0.062674	0.03	rs73033371
19	18329785	G	A	*PDE4C*	0.000581	0.03	rs11879710
22	31018975	T	C	*TCN2*	0.050116	0.03	rs1131603

H1 Cromossoma	Localização	Ref	Alt	Gene	ESP (americanos europeus)	1000 Genomas (população mundial)	dbSNP137
22	35478529	G	A	*ISX*	0.053721	0.03	rs8140287
22	43568512	C	T	*TTLL12*	0.237558	0.20	rs34074034
22	50297888	T	C	*ALG12*	0.114767	0.09	rs3922872

H2 Cromossoma	Localização	Ref	Alt	Gene	ESP (americanos europeus)	1000 Genomas (população mundial)	dbSNP137
1	25572984	T	G	*C1orf63*	0.039884	0.02	rs34484514
1	32145693	C	T	*COL16A1*	0.004866	0.0032	rs41263969
1	38329999	C	G	*INPP5B*	0.143993	0.07	rs41311191
1	43919081	T	C	*HYI*	0.025023	0.01	rs142369206
1	46774783	A	G	*UQCRH*	0.105930	0.05	rs41292543
1	67450431	A	G	*MIER1*	0.026033	0.01	rs17129563
1	64097432	C	T	*PGM1*	0.234419	0.24	rs1126728
1	203186947	G	C	*CHIT1*	NA	NA	rs201682373
1	215960153	A	C	*USH2A*	NA	NA	NA
1	115231254	G	A	*AMPD1*	0.133023	0.05	rs61752479
1	160327026	G	A	*NCSTN*	NA	NA	rs201760425
1	186088994	G	A	*HMCN1*	0.001279	0.0009	rs146671871
1	201017805	A	C	*CACNA1S*	NA	NA	rs79011683
1	208272313	A	C	*PLXNA2*	NA	NA	rs200698765
2	3392295	A	G	*TTC15*	0.374070	0.44	rs11686212
2	74177777	A	G	*DGUOK*	0.022209	0.01	rs74874677
2	71163086	T	C	*ATP6V1B1*	0.436512	0.39	rs11681642
2	220497108	A	G	*SLC4A3*	NA	NA	NA
2	220072702	C	T	*ZFAND2B*	NA	NA	NA
2	143685260	C	T	*KYNU*	NA	NA	NA
2	128477084	G	C	*WDR33*	0.087579	0.05	rs61730415
2	202700399	C	T	*CDK15*	0.022558	0.01	rs34851370
2	230861519	G	T	*FBXO36*	0.240000	0.21	rs1035834
3	67053926	T	C	*KBTBD8*	0.066047	0.03	rs75804175
3	183995137	T	G	*ECE2*	NA	NA	NA
3	56628031	G	A	*CCDC66*	0.537326	0.28	rs7637449
3	12858847	C	T	*CAND2*	0.000238	NA	rs201715663
3	193209178	T	C	*ATP13A4*	0.404186	0.45	rs6788448
3	193171979	T	A	*ATP13A4*	0.106163	0.06	rs35424709
3	52240678	A	G	*ALAS1*	NA	NA	NA
3	165548529	T	C	*BCHE*	0.019884	0.01	rs1799807
3	169569432	C	G	*LRRC31*	0.070611	0.06	rs35923425
4	81952637	T	A	*BMP3*	0.027681	0.01	rs74764079
4	107163667	T	A	*TBCK*	0.003023	0.0014	rs34840340
4	108935600	T	G	*HADH*	0.006512	0.0032	rs61735992
4	159161483	A	T	*TMEM144*	0.009419	0.0041	rs62335898
5	33951693	C	G	*SLC45A2*	0.959186	0.44	rs16891982
5	180659868	G	T	*TRIM41*	NA	NA	NA
5	31799371	G	A	*PDZD2*	0.013488	0.0041	rs116598198
5	37302906	A	G	*NUP155*	NA	NA	NA

H2 Cromossoma	Localização	Ref	Alt	Gene	ESP (americanos europeus)	1000 Genomas (população mundial)	dbSNP137
5	75960968	G	C	*IQGAP2*	0.005000	0.0023	rs34968964
5	102295584	C	A	*PAM*	0.000116	NA	rs148182169
5	120021817	C	A	*PRR16*	0.158488	0.10	rs17853861
6	7565727	A	T	*DSP*	0.041860	0.02	rs17604693
6	126317171	G	A	*TRMT11*	NA	NA	NA
6	155561796	C	T	*TIAM2*	0.289186	0.13	rs11751128
6	112457383	G	C	*LAMA4*	0.260349	0.23	rs1050349
6	116783330	G	A	*FAM26F*	0.186991	0.23	rs1057192
6	117130544	A	C	*GPRC6A*	0.336860	0.18	rs28360548
6	152129063	C	T	*ESR1*	0.005116	0.0014	rs139960913
7	73513555	G	A	*LIMK1*	0.000116	NA	NA
7	87913221	C	T	*STEAP4*	0.199637	0.10	rs34741656
8	79629586	C	T	*FAM164A*	0.020698	0.01	rs77219198
8	101718965	G	A	*PABPC1*	NA	NA	rs62513921
8	101721817	T	C	*PABPC1*	NA	NA	rs201076736
8	101724606	G	A	*PABPC1*	NA	NA	rs202060459
8	59409220	C	A	*CYP7A1*	NA	NA	NA
8	101730043	G	C	*PABPC1*	NA	NA	rs113614781
8	101730073	T	C	*PABPC1*	NA	0.26	rs72681443
9	134955204	T	G	*MED27*	NA	NA	NA
9	140100317	C	G	*NDOR1*	0.115771	0.14	rs113809617
9	129854078	G	A	*ANGPTL2*	NA	NA	NA
9	116082647	C	G	*WDR31*	0.208953	0.10	rs41307479
9	125486717	G	A	*OR1L4*	NA	NA	rs76170289
9	125512575	G	A	*OR1L6*	NA	0.92	rs4838012
9	140100317	C	G	*NDOR1*	0.115771	0.14	rs113809617
10	13699338	T	G	*FRMD4A*	NA	NA	rs199968440
10	26434455	G	T	*MYO3A*	0.093256	0.04	rs33947968
10	72360387	G	A	*PRF1*	0.046279	0.02	rs35947132
10	126682443	G	T	*CTBP2*	NA	NA	rs76582415
10	126683071	G	C	*CTBP2*	NA	NA	rs61870306
10	126683075	T	A	*CTBP2*	NA	NA	rs80273852
10	126683111	C	T	*CTBP2*	NA	NA	rs78155918
10	126683123	A	C	*CTBP2*	NA	NA	rs79936509
10	126683132	A	T	*CTBP2*	NA	NA	rs150320719
10	126683146	C	A	*CTBP2*	NA	NA	rs80025996
10	126683151	C	T	*CTBP2*	NA	NA	rs76203768
10	126683162	C	T	*CTBP2*	NA	NA	rs75420260
10	126683190	C	A	*CTBP2*	NA	NA	rs78681531
10	126686629	G	A	*CTBP2*	NA	NA	rs74523764

H2 Cromossoma	Localização	Ref	Alt	Gene	ESP (americanos europeus)	1000 Genomas (população mundial)	dbSNP137
10	126686724	T	C	*CTBP2*	NA	NA	rs75044667
10	126691575	T	C	*CTBP2*	NA	NA	rs112239066
10	126691660	G	A	*CTBP2*	NA	NA	rs3198935
10	126691926	G	A	*CTBP2*	NA	NA	rs3198932
10	126691937	G	C	*CTBP2*	NA	NA	rs1058301
10	71168766	C	G	*TACR2*	0.001977	0.03	rs79030584
10	126691979	C	G	*CTBP2*	NA	NA	rs3198926
10	126692017	G	C	*CTBP2*	NA	NA	rs3198920
11	26718732	C	A	*SLC5A12*	NA	NA	rs72883299
11	64813384	G	C	*NAALADL1*	0.000815	NA	rs142938335
11	72560874	C	T	*FCHSD2*	NA	NA	NA
11	47660334	C	T	*MTCH2*	NA	NA	NA
11	7818056	T	C	*OR5P2*	0.282502	0.27	rs73406604
11	68707139	T	G	*IGHMBP2*	0.001630	0.0005	rs147674615
11	126146290	T	G	*FOXRED1*	NA	NA	rs201062084
12	2973547	G	T	*FOXM1*	0.020698	0.01	rs28990715
12	104131413	G	T	*STAB2*	NA	NA	NA
12	122091022	G	A	*MORN3*	NA	NA	NA
12	70189021	C	G	*RAB3IP*	NA	NA	NA
12	42499701	C	A	*GXYLT1*	NA	NA	rs74583427
12	42499711	C	A	*GXYLT1*	NA	NA	rs76555438
12	44177511	G	A	*IRAK4*	0.015116	0.01	rs55944915
12	48272895	A	G	*VDR*	0.614070	0.65	rs2228570
12	53343059	C	A	*KRT18*	NA	NA	rs78343594
12	53343209	G	A	*KRT18*	NA	NA	rs79346135
12	58220816	A	G	*CTDSP2*	NA	NA	rs76940645
12	122845698	C	A	*CLIP1*	NA	NA	rs79909185
12	131311749	A	C	*STX2*	0.025465	0.01	rs137928907
14	24707479	G	A	*GMPR2*	0.048357	0.02	rs34354104
14	55448409	G	C	*WDHD1*	0.097326	0.05	rs61741224
14	88935307	C	T	*PTPN21*	NA	NA	NA
14	62229285	T	G	*SNAPC1*	0.006047	0.03	rs74810099
14	81972441	T	C	*SEL1L*	0.014419	0.01	rs11499034
14	90770476	T	A	*C14orf102*	0.003488	0.0009	rs143888178
15	91326099	C	T	*BLM*	0.065263	0.05	rs11852361
15	91491409	C	T	*UNC45A*	0.000233	NA	rs200394198
16	48204078	T	A	*ABCC11*	0.104884	0.05	rs61739606
16	80667116	C	A	*CDYL2*	0.000116	NA	NA
16	84089622	C	T	*MBTPS1*	0.008372	0.0027	rs146299475
17	21204187	G	T	*MAP2K3*	NA	0.50	rs56067280

H2 Cromossoma	Localização	Ref	Alt	Gene	ESP (americanos europeus)	1000 Genomas (população mundial)	dbSNP137
17	21204192	C	T	*MAP2K3*	NA	0.50	rs56216806
17	21207834	C	T	*MAP2K3*	NA	0.30	rs58609466
17	21217513	G	A	*MAP2K3*	NA	0.49	rs2363198
17	21319069	G	A	*KCNJ12*	NA	NA	rs76265595
17	21319079	C	A	*KCNJ12*	NA	NA	rs76518282
17	21319087	G	A	*KCNJ12*	NA	NA	rs75029097
17	21319121	C	T	*KCNJ12*	NA	0.50	rs1714864
17	37243923	G	A	*PLXDC1*	0.000698	NA	rs35062505
17	40824339	G	A	*PLEKHH3*	0.012188	0.0041	rs200210041
17	61903439	T	C	*FTSJ3*	NA	NA	NA
17	46654328	C	A	*HOXB4*	NA	NA	NA
17	45216172	A	G	*CDC27*	NA	NA	rs76836152
17	45235598	G	C	*CDC27*	NA	NA	rs77467652
17	45235635	A	C	*CDC27*	NA	NA	rs79936417
17	45249391	T	G	*CDC27*	NA	NA	rs62077270
17	73520359	C	G	*TSEN54*	0.065465	0.03	rs62088470
17	74162548	C	T	*RNF157*	0.050930	0.02	rs11539879
18	580623	T	C	*CETN1*	0.145814	0.17	rs61734344
18	6975995	G	C	*LAMA1*	0.020349	0.01	rs117225191
18	61654463	A	G	*SERPINB8*	0.580116	0.55	rs3826616
18	3457606	C	T	*TGIF1*	0.077674	0.04	rs4468717
19	39908214	C	T	*PLEKHG2*	0.062674	0.03	rs73033371
19	39915758	C	G	*PLEKHG2*	0.529800	0.56	rs31728
19	4359191	C	T	*MPND*	0.045685	0.02	rs34522164
20	3682126	C	T	*SIGLEC1*	0.070349	0.03	rs34924243
20	43530277	G	T	*YWHAB*	NA	NA	NA
20	6033004	G	A	*LRRN4*	0.164698	0.26	rs6117050
20	44048972	A	C	*PIGT*	0.026512	0.01	rs61753669
22	24179922	G	C	*DERL3*	0.158953	0.15	rs3177243
22	40797647	T	C	*SGSM3*	0.011512	0.0027	rs9611338
22	43568512	C	T	*TTLL12*	0.237558	0.20	rs34074034
22	50297888	T	C	*ALG12*	0.114767	0.09	rs3922872

H3 Cromossoma	Localização	Ref	Alt	Gene	ESP (americanos europeus)	1000 Genomas (população mundial)	dbSNP137
1	29438907	C	T	*EPB41*	0.000581	0.0005	rs148459745
1	34667784	A	C	*C1orf94*	NA	NA	rs80015626
1	40980668	T	C	*DEM1*	0.065698	0.03	rs35672330
1	57422511	C	T	*C8B*	0.055698	0.04	rs12067507
1	1686081	G	A	*NADK*	0.031512	0.02	rs75816936
1	169483561	T	C	*F5*	0.061744	0.05	rs6027
1	179095706	C	T	*ABL2*	NA	NA	NA
2	3392295	A	G	*TTC15*	0.374070	0.44	rs11686212
2	17897488	T	C	*SMC6*	0.264186	0.13	rs35195207
2	28761981	G	C	*PLB1*	0.260930	0.21	rs6753929
2	28825724	A	T	*PLB1*	0.010930	0.01	rs62131028
2	170003432	T	G	*LRP2*	0.790116	0.54	rs4667591
2	107446665	G	C	*ST6GAL2*	0.011744	0.01	rs144829459
2	74177777	A	G	*DGUOK*	0.022209	0.01	rs74874677
2	105859249	C	T	*GPR45*	0.145698	0.10	rs35946826
2	105984191	G	A	*FHL2*	0.000465	NA	rs140148322
2	128046416	G	A	*ERCC3*	0.001628	NA	rs145201970
2	128408389	C	T	*GPR17*	NA	0.0027	rs61738377
2	230861519	G	T	*FBXO36*	0.240000	0.21	rs1035834
3	12962074	G	A	*IQSEC1*	0.135349	0.11	rs35319679
3	15628040	T	A	*HACL1*	0.051524	0.02	rs74637339
3	20153128	C	G	*KAT2B*	NA	NA	NA
3	43618558	C	T	*ANO10*	0.026744	0.01	rs41289586
3	67053926	T	C	*KBTBD8*	0.066047	0.03	rs75804175
3	121155122	C	T	*POLQ*	0.021977	0.01	rs41540016
4	96046223	T	C	*BMPR1B*	NA	NA	NA
4	106859549	G	C	*NPNT*	0.325233	0.32	rs35132891
5	33951693	C	G	*SLC45A2*	0.959186	0.44	rs16891982
5	96513471	G	C	*RIOK2*	0.397442	0.36	rs2544773
5	68695940	T	G	*RAD17*	0.321744	0.26	rs1045051
5	168199842	C	T	*SLIT3*	NA	NA	NA
6	7582993	A	T	*DSP*	0.013023	0.01	rs78652302
6	18143837	T	G	*TPMT*	0.000233	NA	rs151149760
6	21065449	C	T	*CDKAL1*	0.044767	0.06	rs77152992
6	74497102	G	A	*CD109*	0.021512	0.02	rs35466124
6	86195033	G	A	*NT5E*	0.005000	0.0018	rs41271617
6	88123549	G	A	*C6orf165*	0.020843	0.01	rs35438647
6	117130544	A	C	*GPRC6A*	0.336860	0.18	rs28360548
6	138202365	G	A	*TNFAIP3*	0.000349	NA	NA
6	152806014	C	T	*SYNE1*	0.001047	0.0027	rs146366996
7	23728923	C	T	*C7orf46*	0.000698	NA	rs146280373

H3 Cromossoma	Localização	Ref	Alt	Gene	ESP (americanos europeus)	1000 Genomas (população mundial)	dbSNP137
7	36396664	T	C	*KIAA0895*	0.015526	0.01	rs35828246
7	51094269	G	A	*COBL*	0.079186	0.03	rs61737954
7	87913221	C	T	*STEAP4*	0.199637	0.10	rs34741656
7	104747899	G	T	*MLL5*	0.046395	0.02	rs117986340
8	38913228	G	A	*ADAM9*	NA	NA	NA
8	67341714	G	C	*RRS1*	0.008702	0.02	rs34077648
8	101718965	G	A	*PABPC1*	NA	NA	rs62513921
8	101721812	G	A	*PABPC1*	NA	NA	rs200409148
8	101721817	T	C	*PABPC1*	NA	NA	rs201076736
8	101724924	G	A	*PABPC1*	NA	NA	NA
8	101730043	G	C	*PABPC1*	NA	NA	rs113614781
8	101730073	T	C	*PABPC1*	NA	0.26	rs72681443
8	130789767	G	A	*GSDMC*	0.049767	0.02	rs10090835
8	121293179	G	A	*COL14A1*	0.029651	NA	rs117868350
8	146068464	G	A	*ZNF7*	NA	NA	NA
9	108363426	C	T	*FKTN*	0.028282	0.02	rs41277797
9	125512575	G	A	*OR1L6*	NA	0.92	rs4838012
9	131019738	T	C	*GOLGA2*	0.056163	0.14	rs2240961
9	139413097	T	G	*NOTCH1*	NA	NA	rs200520088
10	15145453	C	T	*RPP38*	0.001860	0.0014	rs41284459
10	26508143	C	A	*GAD2*	0.007442	0.0041	rs2839672
10	101180534	C	A	*GOT1*	NA	NA	NA
10	126682484	G	A	*CTBP2*	NA	NA	NA
10	126682485	C	T	*CTBP2*	NA	NA	NA
10	126683071	G	C	*CTBP2*	NA	NA	rs61870306
10	126683075	T	A	*CTBP2*	NA	NA	rs80273852
10	126683123	A	C	*CTBP2*	NA	NA	rs79936509
10	126683132	A	T	*CTBP2*	NA	NA	rs150320719
10	126683146	C	A	*CTBP2*	NA	NA	rs80025996
10	126683151	C	T	*CTBP2*	NA	NA	rs76203768
10	126683162	C	T	*CTBP2*	NA	NA	rs75420260
10	126683219	T	G	*CTBP2*	NA	NA	NA
10	126683225	C	G	*CTBP2*	NA	NA	NA
10	126683235	C	A	*CTBP2*	NA	NA	NA
10	126683243	C	T	*CTBP2*	NA	NA	NA
10	126683244	C	G	*CTBP2*	NA	NA	NA
10	126686593	G	C	*CTBP2*	NA	NA	NA
10	126686629	G	A	*CTBP2*	NA	NA	rs74523764
10	126686692	C	T	*CTBP2*	NA	NA	NA
10	126686724	T	C	*CTBP2*	NA	NA	rs75044667
10	126691575	T	C	*CTBP2*	NA	NA	rs112239066

H3 Cromossoma	Localização	Ref	Alt	Gene	ESP (americanos europeus)	1000 Genomas (população mundial)	dbSNP137
10	126691579	C	T	*CTBP2*	NA	NA	NA
10	126691580	G	C	*CTBP2*	NA	NA	NA
10	126691660	G	A	*CTBP2*	NA	NA	rs3198935
10	126691926	G	A	*CTBP2*	NA	NA	rs3198932
10	126691937	G	C	*CTBP2*	NA	NA	rs1058301
10	126691979	C	G	*CTBP2*	NA	NA	rs3198926
10	126692017	G	C	*CTBP2*	NA	NA	rs3198920
10	126692029	G	T	*CTBP2*	NA	NA	NA
10	126692035	G	T	*CTBP2*	NA	NA	NA
10	126692037	C	A	*CTBP2*	NA	NA	NA
10	95093619	G	A	*MYOF*	0.016845	0.01	rs61861290
10	129690826	A	T	*CLRN3*	0.091977	0.05	rs35070529
11	5068297	G	A	*OR52J3*	0.123197	0.11	rs58664826
11	124294293	T	C	*OR8B4*	NA	NA	NA
11	7818056	T	C	*OR5P2*	0.282502	0.27	rs73406604
11	26718732	C	A	*SLC5A12*	NA	NA	rs72883299
11	121367627	G	A	*SORL1*	0.010235	0.01	rs117260922
12	42499690	T	C	*GXYLT1*	NA	NA	rs79044728
12	42499694	A	T	*GXYLT1*	NA	NA	NA
12	42499701	C	A	*GXYLT1*	NA	NA	rs74583427
12	42499738	T	C	*GXYLT1*	NA	NA	NA
12	42499739	C	T	*GXYLT1*	NA	NA	rs77582546
12	42499825	A	C	*GXYLT1*	NA	NA	rs76034661
12	48272895	A	G	*VDR*	0.614070	0.65	rs2228570
12	50367201	A	C	*AQP6*	NA	NA	rs202163475
12	58220816	A	G	*CTDSP2*	NA	NA	rs76940645
12	70091452	A	G	*BEST3*	0.088156	0.04	rs1025016
12	100708367	C	T	*SCYL2*	0.119386	0.05	rs33968174
12	109665242	G	A	*ACACB*	0.060698	0.05	rs60293430
12	101693534	C	G	*UTP20*	0.189230	0.11	rs4764643
12	122845698	C	A	*CLIP1*	NA	NA	rs79909185
13	21751164	C	T	*MRP63*	0.011520	0.0046	rs117575190
14	51382183	C	T	*PYGL*	0.000814	NA	rs2228499
14	54997751	C	T	*CGRRF1*	0.010468	0.01	rs34839928
14	81737076	A	C	*STON2*	0.589535	0.44	rs2241621
14	62229285	T	G	*SNAPC1*	0.006047	0.03	rs74810099
14	75747512	G	C	*FOS*	0.006164	0.01	rs138334429
14	103342015	C	T	*TRAF3*	0.001744	0.0005	rs143813189
15	40398285	C	A	*BMF*	0.000116	NA	NA
15	79224727	T	C	*CTSH*	0.026089	0.01	rs78155742
16	420140	G	A	*MRPL28*	0.169695	0.10	rs3194151

H3 Cromossoma	Localização	Ref	Alt	Gene	ESP (americanos europeus)	1000 Genomas (população mundial)	dbSNP137
16	11785220	G	A	*TXNDC11*	0.001395	0.0018	rs17674449
16	19049346	T	C	*TMC7*	0.000233	0	rs148734728
16	81232275	G	A	*PKD1L2*	0.792651	0.80	rs7205673
16	65022234	C	T	*CDH11*	0.119767	0.22	rs1130821
16	67697922	C	T	*C16orf48*	0.025752	0.01	rs143645135
16	67970188	G	C	*PSMB10*	NA	NA	NA
16	71683718	A	G	*PHLPP2*	0.149884	0.10	rs61733127
17	21204187	G	T	*MAP2K3*	NA	0.50	rs56067280
17	21204192	C	T	*MAP2K3*	NA	0.50	rs56216806
17	21207834	C	T	*MAP2K3*	NA	0.30	rs58609466
17	21217513	G	A	*MAP2K3*	NA	0.49	rs2363198
17	21319069	G	A	*KCNJ12*	NA	NA	rs76265595
17	21319079	C	A	*KCNJ12*	NA	NA	rs76518282
17	21319087	G	A	*KCNJ12*	NA	NA	rs75029097
17	21319121	C	T	*KCNJ12*	NA	0.50	rs1714864
17	33497222	A	T	*UNC45B*	0.002326	0.0005	rs142883160
17	42426836	T	C	*GRN*	NA	NA	NA
17	45216172	A	G	*CDC27*	NA	NA	rs76836152
17	45235598	G	C	*CDC27*	NA	NA	rs77467652
17	45235635	A	C	*CDC27*	NA	NA	rs79936417
17	45249391	T	G	*CDC27*	NA	NA	rs62077270
17	47284735	T	C	*GNGT2*	0.054535	0.03	rs35638197
17	64210580	A	C	*APOH*	0.036395	0.02	rs1801689
17	72758167	G	A	*SLC9A3R1*	0.004535	0.0009	NA
17	11651057	A	G	*DNAH9*	0.244070	0.23	rs3744581
17	74162548	C	T	*RNF157*	0.050930	0.02	rs11539879
18	580623	T	C	*CETN1*	0.145814	0.17	rs61734344
18	72228211	G	T	*CNDP1*	NA	NA	NA
19	7965119	G	A	*LRRC8E*	NA	NA	NA
19	7998386	G	C	*TIMM44*	0.040581	0.01	rs118048213
19	44057574	G	A	*XRCC1*	0.064535	0.13	rs1799782
19	48377970	T	A	*SULT2A1*	0.000465	NA	rs148329743
19	1036156	G	A	*CNN2*	NA	NA	NA
19	1036157	T	G	*CNN2*	NA	NA	NA
19	1036165	G	A	*CNN2*	NA	NA	NA
19	1036169	T	A	*CNN2*	NA	NA	NA
19	1037715	T	C	*CNN2*	NA	NA	rs200177867
19	1037716	G	A	*CNN2*	NA	NA	rs201532581
19	1037718	G	T	*CNN2*	NA	NA	rs199741851
19	1037756	G	A	*CNN2*	NA	NA	rs77830704
19	1037766	G	A	*CNN2*	NA	NA	rs78386506

H3 Cromossoma	Localização	Ref	Alt	Gene	ESP (americanos europeus)	1000 Genomas (população mundial)	dbSNP137
19	1037781	C	A	*CNN2*	NA	NA	rs75676484
20	3180666	C	A	*DDRGK1*	0.026860	0.01	rs35327491
20	3682126	C	T	*SIGLEC1*	0.070349	0.03	rs34924243
20	6033004	G	A	*LRRN4*	0.164698	0.26	rs6117050
21	34861220	G	A	*DNAJC28*	NA	NA	NA
22	30864655	A	G	*SEC14L3*	0.016628	0.01	rs115278158
22	30866053	G	A	*SEC14L3*	0.016744	0.01	rs114438349
22	40797647	T	C	*SGSM3*	0.011512	0.0027	rs9611338
22	47106987	C	T	*CERK*	NA	NA	NA
22	50297888	T	C	*ALG12*	0.114767	0.09	rs3922872
X	129518607	G	T	*GPR119*	NA	NA	NA

H4 Cromossoma	Localização	Ref	Alt	Gene	ESP (americanos europeus)	1000 Genomas (população mundial)	dbSNP137
1	3624280	C	A	*TP73*	NA	NA	NA
1	16456763	C	T	*EPHA2*	0.025581	0.01	rs35903225
1	17588689	T	A	*PADI3*	0.007209	0.0037	rs142129409
1	18961000	C	T	*PAX7*	0.000116	NA	rs148832029
1	34667784	A	C	*C1orf94*	NA	NA	rs80015626
1	37285454	A	C	*GRIK3*	NA	NA	rs199790296
1	115231254	G	A	*AMPD1*	0.133023	0.05	rs61752479
1	119427467	A	C	*TBX15*	0.047791	0.03	rs61730011
1	201017805	A	C	*CACNA1S*	NA	NA	rs79011683
1	156526387	C	G	*IQGAP3*	0.638372	0.47	rs11264498
1	230907799	C	T	*CAPN9*	0.001977	0.0005	rs28359655
2	3392295	A	G	*TTC15*	0.374070	0.44	rs11686212
2	26818069	C	A	*CIB4*	0.008023	0.0023	rs112421794
2	105859249	C	T	*GPR45*	0.145698	0.10	rs35946826
2	148676144	A	C	*ACVR2A*	NA	NA	NA
2	160114355	A	G	*WDSUB1*	0.000116	NA	rs142150362
2	183070764	G	C	*PDE1A*	NA	NA	NA
2	216288103	C	T	*FN1*	NA	NA	NA
2	228163475	G	A	*COL4A3*	0.000243	0.0014	rs190598500
2	230861519	G	T	*FBXO36*	0.240000	0.21	rs1035834
2	241808314	C	T	*AGXT*	0.203419	0.11	rs34116584
3	52620701	A	T	*PBRM1*	NA	NA	NA
3	133534466	T	C	*SRPRB*	NA	0.02	rs113131025
4	100534222	G	C	*MTTP*	NA	NA	NA
5	68695940	T	G	*RAD17*	0.321744	0.26	rs1045051
5	86695274	A	G	*CCNH*	0.223837	0.14	rs2230641
5	145895520	C	T	*GPR151*	NA	NA	NA
5	145895558	G	A	*GPR151*	0.075698	0.04	rs17104742
5	112179572	C	T	*APC*	NA	NA	NA
6	42196127	C	T	*TRERF1*	0.051279	0.03	rs11751765
6	52701085	T	C	*GSTA5*	0.004651	0.0018	rs146408369
6	152772264	A	G	*SYNE1*	0.429535	0.59	rs214976
6	29408090	A	G	*OR10C1*	0.017165	0.06	rs17177632
6	135239816	G	T	*ALDH8A1*	0.005930	0.0018	rs61731731
6	160575837	G	A	*SLC22A1*	0.026047	0.01	rs34059508
7	56087374	C	T	*PSPH*	NA	NA	rs200442078
8	41577289	C	T	*ANK1*	0.001977	0.0009	rs147608206
8	101718965	G	A	*PABPC1*	NA	NA	rs62513921
8	101724606	G	A	*PABPC1*	NA	NA	rs202060459
8	101724924	G	A	*PABPC1*	NA	NA	NA

H4 Cromossoma	Localização	Ref	Alt	Gene	ESP (americanos europeus)	1000 Genomas (população mundial)	dbSNP137
8	101730043	G	C	*PABPC1*	NA	NA	rs113614781
8	101730073	T	C	*PABPC1*	NA	0.26	rs72681443
9	368128	C	T	*DOCK8*	0.106047	0.04	rs17673268
9	104433234	C	A	*GRIN3A*	0.001395	0.0005	rs139560863
9	117170241	G	C	*DFNB31*	0.122005	0.07	rs12339210
9	125486717	G	A	*OR1L4*	NA	NA	rs76170289
9	125512575	G	A	*OR1L6*	NA	0.92	rs4838012
9	139413097	T	G	*NOTCH1*	NA	NA	rs200520088
9	139835767	C	T	*FBXW5*	NA	NA	NA
9	140100317	C	G	*NDOR1*	0.115771	0.14	rs113809617
9	140130521	C	T	*SLC34A3*	0.000466	NA	rs145029982
10	32575886	A	G	*EPC1*	NA	NA	NA
10	29754535	G	A	*SVIL*	0.130961	0.06	rs17694739
10	29784072	G	C	*SVIL*	NA	0.64	rs2368406
10	47087501	C	T	*PPYR1*	0.284884	0.30	rs3824733
10	113921450	C	A	*GPAM*	NA	NA	rs199856746
10	74886544	A	G	*NUDT13*	0.016860	0.01	rs17658872
10	92631765	G	T	*RPP30*	NA	NA	NA
10	126686692	C	T	*CTBP2*	NA	NA	NA
10	126691552	T	G	*CTBP2*	NA	NA	NA
10	126691575	T	C	*CTBP2*	NA	NA	rs112239066
10	126691579	C	T	*CTBP2*	NA	NA	NA
10	126691660	G	A	*CTBP2*	NA	NA	rs3198935
10	126691937	G	C	*CTBP2*	NA	NA	rs1058301
10	126691979	C	G	*CTBP2*	NA	NA	rs3198926
10	126692017	G	C	*CTBP2*	NA	NA	rs3198920
10	126692029	G	T	*CTBP2*	NA	NA	NA
10	126692035	G	T	*CTBP2*	NA	NA	NA
10	126692037	C	A	*CTBP2*	NA	NA	NA
11	3681483	C	G	*ART1*	0.008376	0.0018	rs150574054
11	7672920	G	A	*PPFIBP2*	NA	NA	NA
11	9043479	C	G	*SCUBE2*	NA	NA	NA
11	26718732	C	A	*SLC5A12*	NA	NA	rs72883299
11	67842212	G	A	*CHKA*	NA	NA	NA
11	7818056	T	C	*OR5P2*	0.282502	0.27	rs73406604
11	64666182	C	G	*ATG2A*	NA	NA	rs201703555
11	82976978	A	G	*CCDC90B*	0.001978	0.0014	rs144139155
12	42499690	T	C	*GXYLT1*	NA	NA	rs79044728
12	42499694	A	T	*GXYLT1*	NA	NA	NA
12	42499701	C	A	*GXYLT1*	NA	NA	rs74583427
12	42499711	C	A	*GXYLT1*	NA	NA	rs76555438

H4 Cromossoma	Localização	Ref	Alt	Gene	ESP (americanos europeus)	1000 Genomas (população mundial)	dbSNP137
12	42499738	T	C	*GXYLT1*	NA	NA	NA
12	42499739	C	T	*GXYLT1*	NA	NA	rs77582546
12	42499825	A	C	*GXYLT1*	NA	NA	rs76034661
12	42512830	T	A	*GXYLT1*	NA	NA	rs76740071
12	42512910	T	A	*GXYLT1*	NA	NA	rs137952794
12	48272895	A	G	*VDR*	0.614070	0.65	rs2228570
12	50561023	G	A	*LASS5*	0.005682	0.0032	rs143484198
12	52822136	C	G	*KRT75*	NA	NA	NA
12	52824351	C	T	*KRT75*	0.012442	0.0037	rs2232398
12	54394264	G	A	*HOXC9*	NA	NA	NA
12	56494998	A	T	*ERBB3*	0.106512	0.06	rs773123
12	58217453	G	A	*CTDSP2*	NA	NA	NA
12	101693534	C	G	*UTP20*	0.189230	0.11	rs4764643
12	111886074	C	T	*SH2B3*	0.000581	NA	rs72650662
12	117465906	C	T	*FBXW8*	0.000116	0.0005	rs191628514
12	122845698	C	A	*CLIP1*	NA	NA	rs79909185
13	42385446	T	C	*KIAA0564*	0.042674	0.07	rs9562353
13	49775997	A	G	*FNDC3A*	0.070016	0.03	rs45604939
13	99083321	C	T	*FARP1*	0.043256	0.02	rs61730892
13	111164285	C	T	*COL4A2*	NA	NA	NA
13	115047496	G	C	*UPF3A*	NA	NA	rs76186578
14	30046460	C	T	*PRKD1*	0.000814	NA	rs150599710
14	77769235	A	G	*POMT2*	NA	NA	NA
15	69709792	C	T	*KIF23*	0.070149	0.03	rs61751119
15	72645446	C	T	*HEXA*	NA	NA	NA
15	75650822	T	C	*MAN2C1*	0.000233	0.0014	rs143005170
16	420140	G	A	*MRPL28*	0.169695	0.10	rs3194151
16	3721773	G	C	*TRAP1*	0.184070	0.11	rs1136948
16	22328486	C	T	*POLR3E*	0.011395	0.01	rs143440180
16	25255503	G	C	*ZKSCAN2*	0.000814	0.0005	rs116889865
16	29908385	G	A	*SEZ6L2*	0.003042	0.0037	rs184056093
16	57931428	C	T	*CNGB1*	0.005803	0.0023	rs148999583
16	67970188	G	C	*PSMB10*	NA	NA	NA
16	65022234	C	T	*CDH11*	0.119767	0.22	rs1130821
16	21051209	G	C	*DNAH3*	0.100930	0.15	rs330150
16	81211496	C	A	*PKD1L2*	0.168336	0.17	rs9935113
16	70563087	C	G	*SF3B3*	NA	NA	rs76371422
17	17720319	C	T	*SREBF1*	0.015465	0.01	rs36215896
17	21204187	G	T	*MAP2K3*	NA	0.50	rs56067280
17	21204192	C	T	*MAP2K3*	NA	0.50	rs56216806
17	21207834	C	T	*MAP2K3*	NA	0.30	rs58609466

H4 Cromossoma	Localização	Ref	Alt	Gene	ESP (americanos europeus)	1000 Genomas (população mundial)	dbSNP137
17	21217513	G	A	*MAP2K3*	NA	0.49	rs2363198
17	21319069	G	A	*KCNJ12*	NA	NA	rs76265595
17	21319079	C	A	*KCNJ12*	NA	NA	rs76518282
17	21319087	G	A	*KCNJ12*	NA	NA	rs75029097
17	21319121	C	T	*KCNJ12*	NA	0.50	rs1714864
17	40824339	G	A	*PLEKHH3*	0.012188	0.0041	rs200210041
17	45216172	A	G	*CDC27*	NA	NA	rs76836152
17	45232043	T	C	*CDC27*	NA	NA	rs186452221
17	45235598	G	C	*CDC27*	NA	NA	rs77467652
17	45235635	A	C	*CDC27*	NA	NA	rs79936417
17	45249391	T	G	*CDC27*	NA	NA	rs62077270
17	48761053	G	A	*ABCC3*	0.063023	0.03	rs11568591
17	56598991	T	A	*SEPT4*	0.120581	0.05	rs17741424
17	55182878	C	T	*AKAP1*	0.102209	0.06	rs17761023
17	68128332	G	A	*KCNJ16*	0.000814	NA	rs146940841
18	43796258	C	T	*C18orf25*	NA	NA	NA
19	3752601	G	A	*APBA3*	0.019277	0.01	rs147130540
19	7697302	C	T	*PCP2*	0.000116	NA	rs142018557
19	7708058	C	T	*STXBP2*	0.016163	0.01	rs117761837
19	7964582	G	A	*LRRC8E*	NA	NA	NA
19	17837593	C	G	*MAP1S*	NA	NA	rs201564716
20	44516175	C	T	*C20orf165*	NA	NA	NA
20	57599303	C	T	*TUBB1*	0.019186	0.01	rs35565630
20	60776027	C	T	*GTPBP5*	0.002442	0.0023	rs143916641
21	35239562	A	G	*ITSN1*	0.073721	0.06	rs56279221
21	37660314	C	T	*DOPEY2*	0.002907	0.0014	rs145488940
21	47532381	C	T	*COL6A2*	NA	NA	NA
X	17745633	C	T	*NHS*	NA	NA	NA
X	152936568	A	G	*PNCK*	NA	NA	NA

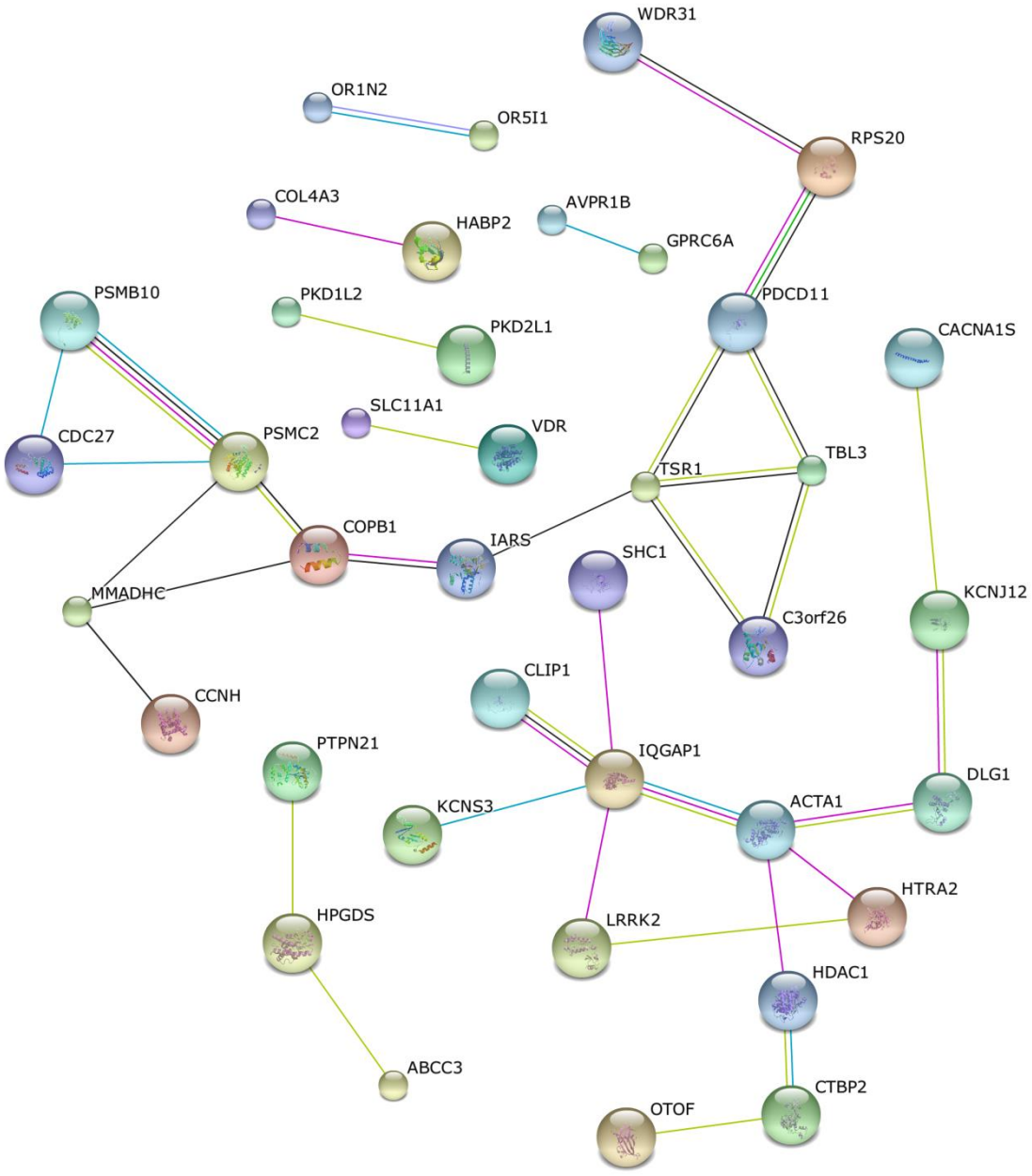
*NA = Não avaliado

Anexo 6: Variantes filtradas pelo SIFT, PolyPhen 2, LRT e *Mutation Taster*, comuns aos quatro genomas estudados.

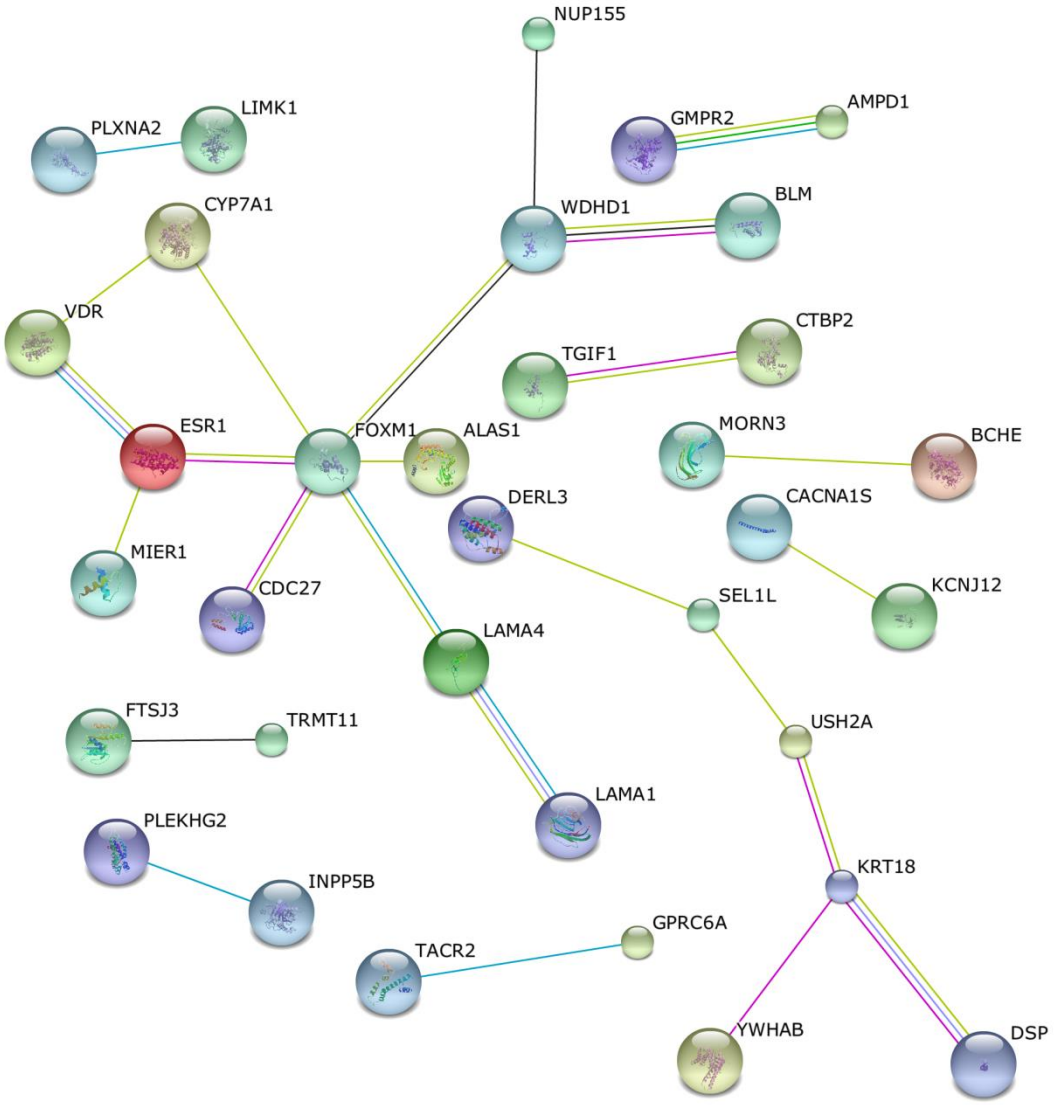
Cromossoma	Localização	Ref	Alt	Gene	ESP (americanos europeus)	1000 Genomas (população mundial)	dbSNP137
2	3392295	A	G	*TRAPPC12*	0.374070	0.44	rs11686212
2	230861519	G	T	*FBXO36*	0.240000	0.21	rs1035834
8	101718965	G	A	*PABPC1*	NA	NA	rs62513921
8	101724924	G	A	*PABPC1*	NA	NA	NA
8	101730043	G	C	*PABPC1*	NA	NA	rs113614781
8	101730073	T	C	*PABPC1*	NA	0.26	rs72681443
9	125512575	G	A	*OR1L6*	NA	0.92	rs4838012
10	126691575	T	C	*CTBP2*	NA	NA	rs112239066
10	126691579	C	T	*CTBP2*	NA	NA	NA
10	126691979	C	G	*CTBP2*	NA	NA	rs3198926
10	126692017	G	C	*CTBP2*	NA	NA	rs3198920
10	126692029	G	T	*CTBP2*	NA	NA	NA
10	126692035	G	T	*CTBP2*	NA	NA	NA
10	126692037	C	A	*CTBP2*	NA	NA	NA
12	48272895	A	G	*VDR*	0.614070	0.65	rs2228570
12	58217453	G	A	*CTDSP2*	NA	NA	NA
12	122845698	C	A	*CLIP1*	NA	NA	rs79909185
17	21204187	G	T	*MAP2K3*	NA	0.50	rs56067280
17	21204192	C	T	*MAP2K3*	NA	0.50	rs56216806
17	21217513	G	A	*MAP2K3*	NA	0.49	rs2363198
17	21319069	G	A	*KCNJ12*	NA	NA	rs76265595
17	21319079	C	A	*KCNJ12*	NA	NA	rs76518282
17	21319087	G	A	*KCNJ12*	NA	NA	rs75029097
17	21319121	C	T	*KCNJ12*	NA	0.50	rs1714864
17	45216172	A	G	*CDC27*	NA	NA	rs76836152
17	45235598	G	C	*CDC27*	NA	NA	rs77467652
17	45235635	A	C	*CDC27*	NA	NA	rs79936417
17	45249391	T	G	*CDC27*	NA	NA	rs62077270

*NA = Não avaliado

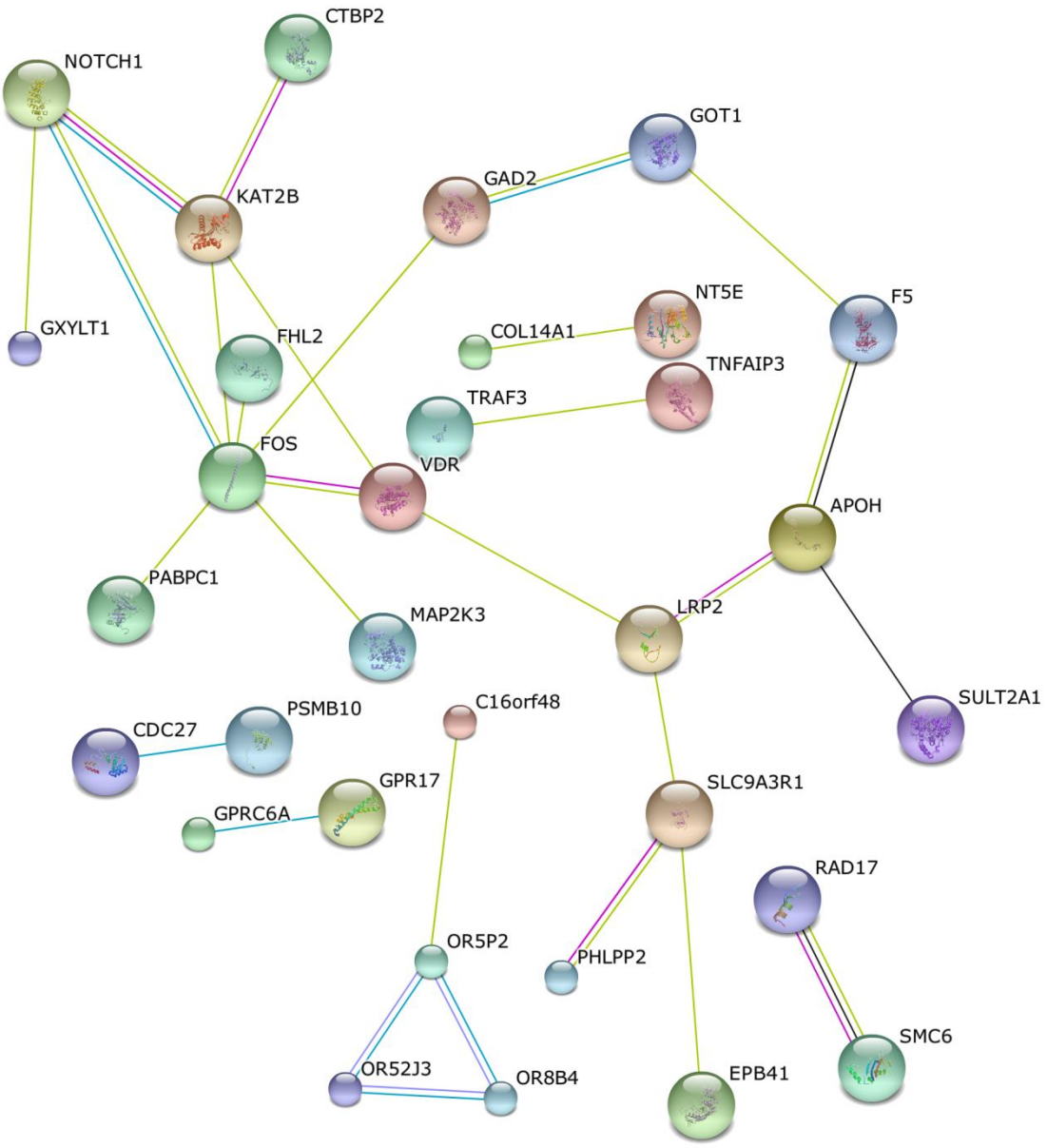
Anexo 7: Análise das interações proteína-proteína do genoma H1, analisadas pelo STRING, onde se observam 36 interações



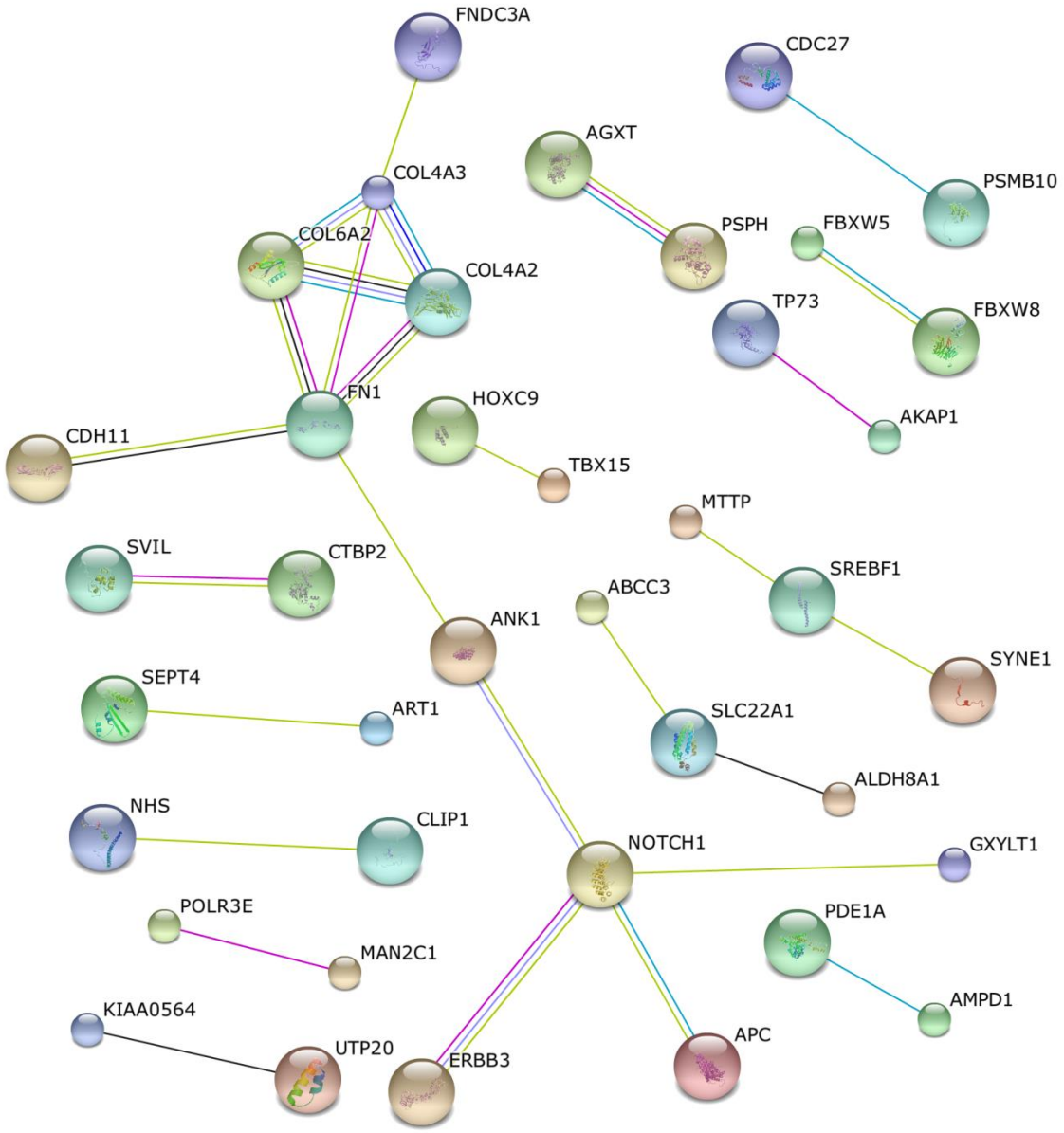
Anexo 8: Análise do total das interações proteína-proteína do genoma H2, onde se observam 25 interações, analisadas pelo STRING.



Anexo 9: Análise do total das interações proteína-proteína do genoma H3, onde se observam 29 interações, analisadas pelo STRING.



Anexo 10: Análise do total das interações proteína-proteína do genoma H4, onde se observam 28 interações, analisadas pelo STRING.



Anexo 11: Função biológica e função molecular dos genes onde foram identificadas variantes não sinónimas, nos genomas H1, H2, H3 e H4.

Gene	Função Molecular	Processo Biológico
ABCC11	Atividade de ATPase Transportador transmembranar	Sistema imune Processo metabólico Resposta a substâncias tóxicas Transporte extracelular
ABCC12	Atividade ATPase Atividade de transportador transmembranar	Sistema imune Processo metabólico Resposta a substâncias tóxicas Transporte extracelular
ABCC3	Atividade ATPase Atividade de transportador transmembranar	Sistema imune Processo metabólico Resposta a substâncias tóxicas Transporte extracelular
ABL2	Atividade tirosina-cinase	Fosforilação proteica Sinalização celular
ACAA1	Atividade de acetiltransferase	Acetilação de proteínas
ACACB	Atividade ligase	Gluconeogénese Biossíntese dos ácidos gordos
ACTA1	Constituinte estrutural do citoesqueleto	Citocinese Mitose Morfogénese Transporte proteico intracelular Exocitose e endocitose Organização dos componentes celulares
ACVR2A	Fator de crescimento Recetor de citosina	Fosforilação proteica Comunicação celular Desenvolvimento da mesoderme Desenvolvimento do sistema esquelético
ADAM9	Atividade proteolítica Atividade de metaloendopeptidase	Processo apoptótico Sistema neurológico Desenvolvimento da mesoderme Processo apoptótico Desenvolvimento cardíaco
ADPRHL1	Atividade de hidrólase Ligação proteica Atividade reguladora GTPase	Processo metabólico e celular Morfogénese Transporte proteico intracelular Exocitose Regulação da atividade catalítica Organização dos componentes celulares
AGXT	Atividade de transaminase	Biossíntese de aminoácidos
AKAP1	Ligação ao RNA	Coagulação sanguínea
ALAS1	Atividade de transferase	Processo metabólico Metabolismo lipídico
ALDH8A1	Atividade de oxireductase	Metabolismo de aminoácidos
ALG12	Atividade de transferase	Tradução Glicosilação proteica
AMPD1	Atividade de hidrólase Atividade de desaminase	Metabolismo das purinas
ANGPTL2	Ligação a recetores	Comunicação celular Adesão

Gene	Função Molecular	Processo Biológico
ANGPTL4	Ligação a recetores	Comunicação celular Adesão Angiogénese
ANK1	Constituinte estrutural do citoesqueleto Ligação à actina	Morfogénese Organização dos componentes celulares
ANK3	Constituinte estrutural do citoesqueleto	Processo celular Morfogénese Organização dos componentes celulares
ANO10	Transporte transmembranar	Morte celular Transporte iónico
APBA3	Ligação à β-amilóide Atividade enzimática	Transporte de proteínas intracelulares Regulação negativa da atividade catalítica
APC	Ligação a cateninas Ligação a microtúbulos	Adesão celular Migração celular Processo apoptótico
APOH	Atividade de peptidase tipo-serina Atividade de metalopeptidase Atividade de recetor Transporte lipídico	Ativação do complemento Proteólise Comunicação celular Adesão Coagulação sanguíneo Transporte lipídico
AQP6	Transportador transmembranar	Transporte
ARRDC4	-	-
ART1	Atividade de transferase	Resposta imune inata
ATG2A	Ligação proteica	Autofagia
ATP13A4	Atividade de hidrólase Canal iónico Transportador transmembranar de catiões	Processo metabólico Transporte de catiões Homeostasia celular do ião cálcio
ATP6V0A2	Atividade de hidrólase Transportador transmembranar de catiões	Sistema imune Processo metabólico Defesa celular Transporte de catiões
ATP6V1B1	Atividade de hidrólase Atividade recetora Canal iónico Transportador transmembranar de catiões	Cadeia respiratória transportadora de eletrões Metabolismo das purinas Transporte de catiões
AVPR1B	Atividade de recetor	Processo celular Contração muscular Sistema neurológico Desenvolvimento
B4GALT4	Atividade de transferase	Processo metabólico de hidratos de carbono Processo metabólico de glicosaminoglicanos
BANK1	-	Ativação de células B
BCHE	Atividade de hidrolase	Processo metabólico
BEST3	Canal iónico	Transporte de aniões
BICD2	Ligação proteica	Morte celular

Gene	Função Molecular	Processo Biológico
BLM	Atividade DNA helicase Ligação a ácidos nucleicos	Replicação de DNA Reparação de DNA Recombinação de DNA Ciclo celular
BMF	Ligação proteica	Processo apoptótico
BMP3	Fator de crescimento	Formação de gâmetas femininos Comunicação celular Desenvolvimento da mesoderme e ectoderme Desenvolvimento do sistema esquelético Desenvolvimento cardíaco Crescimento muscular
BMPR1B	Fator de crescimento Atividade de recetor	Fosforilação proteica Comunicação celular Desenvolvimento da mesoderme Desenvolvimento do sistema esquelético
C11orf66	Ligação de fosfatase	-
C14orf149	Atividade de liase	Processo metabólico
C15orf39	-	-
C16orf48	-	-
C18orf25	-	-
C1orf63	Atividade catalítica Ligação ao mRNA	RNA splicing mRNA splicing via spliceossoma
C1orf94	Ligação proteica	-
C20orf165	-	Diferenciação celular Espermatogénese
C2CD2L	-	-
C3orf20	-	-
C3orf26	Ligação ao RNA poli (A)	-
C6orf165	-	-
C6orf201	-	-
C6orf27	-	-
C7orf46	-	-
C8B	Atividade de peptidase tipo-serina Atividade de metalopeptidase Atividade de recetor Transportador lipídico	Ativação do complemento Proteólise Comunicação celular Adesão Coagulação sanguínea Transporte lipídico
CACNA1S	Canal iónico dependente de voltagem Transporte transmembranar	Transmissão sináptica Contração muscular Secreção de neurotransmissores Propagação do potencial de ação neuronal Transporte de catiões
CAND2	Fator de transcrição	Transcrição do promotor da RNA polimerase II
CAPN9	Atividade de peptidase tipo-cisteína Ligação ao ião cálcio Ligação à calmodulina	Sistema imune Indução da apoptose Proteólise Comunicação celular

Gene	Função Molecular	Processo Biológico
CCBE1	Atividade de recetor Constituinte estrutural da matriz extracelular Transportador transmembranar	Ativação de macrófagos Comunicação celular Adesão Circulação sanguínea Desenvolvimento da ectoderme e mesoderme Morfogénese Resposta a estímulos Transporte proteico intracelular Endocitose Regulação da tensão superficial Organização dos componentes celulares
CCDC66	-	Desenvolvimento da retina
CCDC90B	-	-
CCNH	Fator de transcrição Atividade de cinase Ligação proteica Regulador de cinase	Transcrição mRNA *splicing* via spliceossoma Ciclo celular Regulação da transcrição Regulação da atividade catalítica
CD109	Atividade de peptidase Atividade de citocinas Inibidor de endopeptidase tipo-serina	Ativação do complemento Proteólise Processo celular Resposta a estímulos Regulação da atividade catalítica
CDC27	Ligação proteica	Proliferação celular Ciclo celular Regulação do ciclo celular
CDH11	Ligação	Processo celular Perceção visual e sensorial Desenvolvimento da mesoderme Desenvolvimento do sistema nervoso Desenvolvimento muscular
CDK15	Atividade de tirosina-cinase	Processo metabólico do glicogénio Fosforilação proteica Mitose Comunicação celular Desenvolvimento da ectoderme e mesoderme Desenvolvimento do sistema nervoso
CDKAL1	Atividade de transferase	Manutenção da fidelidade da tradução
CDYL2	Atividade oxireductase Atividade de acetiltransferase Atividade de ligase	Coenzima Biossíntese de vitaminas Metabolismo de hidratos de carbono β-oxidação dos ácidos gordos
CEP128	Atividade de hidrolase Ligação ao ácido nucleico Ligação à cromatina	Replicação de DNA Reparação de DNA Mitose Meiose Segregação de cromossomas Organização da cromatina
CERK	Atividade de cinase	Processo metabólico Comunicação celular
CETN1	Ligação ião cálcio Ligação à calmodulina	Movimento dos componentes celulares Ciclo celular Comunicação celular
CFTR	Atividade ATPase Canal aniónico	Sistema Imune Processo metabólico Resposta a substâncias tóxicas Transporte extracelular
CGRRF1	Ligação a iões de zinco	Resposta ao *stress* Paragem do ciclo celular
CHIT1	Atividade de hidrolase	Metabolismo dos polissacarídeos

Gene	Função Molecular	Processo Biológico
CHKA	Atividade cinase	Processo metabólico de fosfolípidos
CIB4	Atividade de fosfatase Ligação ao ião cálcio Ligação à calmodulina	Modificação proteica Comunicação celular
CLIP1	Constituinte estrutural do citoesqueleto Ligação a microtúbulos	*Folding* das proteínas Processo Celular Morfogénese Transporte intracelular Transporte mediado por vesículas Organização dos componentes celulares
CLRN3	-	Vesicula extracelular
CLSPN	Ligação ao DNA	Reparação de DNA *Checkpoint* G2 do ciclo celular Processo apoptótico
CMPK1	Atividade de nucleótido cinase	Metabolismo das purinas e pirimidinas
CNGB1	Canal iónico dependente de voltagem Transporte transmembranar Ligação	Comunicação celular Resposta a estímulos externos Resposta a estímulos abióticos Transporte de catiões Regulação dos processos biológico
CNN2	Constituinte estrutural do citoesqueleto Ligação à actina	Contração muscular
COBL	-	Desenvolvimento da ectoderme Desenvolvimento do sistema nervoso
COL14A1	Atividade de recetor	Movimento dos componentes celulares Adesão Coagulação sanguínea
COL16A1	Atividade de recetor Constituinte estrutural da matriz extracelular Transportador transmembranar	Ativação de macrófago Comunicação celular Adesão Circulação sanguínea Desenvolvimento da mesoderme e ectoderme Morfogénese Resposta a estímulos Transporte de proteínas intracelulares Endocitose Organização dos componentes celulares
COL4A2	Atividade de recetor Constituinte estrutural da matriz extracelular Transporte transmembranar	Ativação de macrófago Comunicação celular Adesão Circulação sanguínea Desenvolvimento da ectoderme e mesoderme Morfogénese Resposta a estímulos Transporte de proteínas intracelulares Endocitose Organização dos componentes celulares
COL4A3	Atividade de recetor Constituinte estrutural da matriz extracelular Transportador transmembranar	Ativação dos macrófagos Comunicação celular Adesão Circulação sanguínea Desenvolvimento da ectoderme e mesoderme

Gene	Função Molecular	Processo Biológico
COL6A2	Atividade de recetor Constituinte estrutural da matriz extracelular Transporte transmembranar	Ativação de macrófago Comunicação celular Adesão Circulação sanguínea Desenvolvimento da ectoderme e mesoderme Morfogénese Resposta a estímulos Transporte de proteínas intracelulares Endocitose Organização dos componentes celulares
COPB1	Ligação proteica Atividade estrutural da molécula	Transporte intracelular exocitose Organização da membrana celular
CPA5	Atividade de metalopeptidase	Proteólise
CTBP2	Co-fator da transcrição Atividade oxireductase Ligação proteica	Metabolismo de hidratos de carbono Biossíntese de aminoácidos
CTDSP2	Atividade de fosfatase Ligação proteica	Desfosforilação proteica
CTSH	Ligação à hormona tiroideia Atividade aminopeptidase	Citotoxicidade mediada por células T Resposta imune adaptativa Resposta celular ao estímulo da hormona tiroideia
CYP7A1	Atividade oxireductase	Cadeia respiratória transportadora de eletrões Metabolismo do colesterol
DDRGK1	Ligação proteica	-
DEM1	-	Processo catabólico
DERL3	Atividade de recetor	-
DFNB31	Constituinte estrutural do citoesqueleto	Perceção sensorial e sensorial
DGUOK	Atividade de cinase	Processo metabólico de purinas e pirimidinas
DHTKD1	-	-
DLG1	-	Transmissão sináptica Adesão Sistema neurológico Desenvolvimento da ectoderme Morfogénese Sistema nervoso Localização proteica Organização dos componentes celulares
DMRT2	Ligação aos ácidos nucleicos	Fator de transcrição Desenvolvimento da mesoderme Regulação da transcrição
DMXL1	-	Transporte proteico intracelular Exocitose
DNAH3	Atividade motora de microtúbulos Constituinte estrutural do citoesqueleto	Processo metabólico Movimento dos componentes celulares Mitose Segregação de cromossomas Transporte de proteínas intracelulares Transporte mediado por vesículas Organização dos componentes celulares

Gene	Função Molecular	Processo Biológico
DNAH9	Atividade motora de microtúbulos Constituinte estrutural do citoesqueleto	Processo metabólico Movimento dos componentes celulares Mitose Segregação de cromossomas Transporte de proteínas intracelulares Transporte mediado por vesículas Organização dos componentes celulares
DNAJC28	Transportador transmembranar Ligação ao RNA	Transporte de proteínas intracelulares
DNASE1L3	-	-
DOCK8	Atividade catalítica Ligação proteica Atividade regulador GTPase	Processo metabólico Movimento dos componentes celulares Comunicação celular Transporte de proteínas intracelulares Fagocitose Regulação da atividade catalítica
DOPEY2	-	-
DPEP1	Atividade de metalopeptidase	Proteólise
DSP	Constituinte estrutural do citoesqueleto Ligação à actina	Movimento dos componentes celulares Morfogénese Organização dos componentes celulares
ECE2	Atividade metalopeptidase	Proteólise Comunicação celular Regulação do processo biológico
ECEL1	Atividade de metalopeptidase	Proteólise Comunicação celular Circulação sanguínea Regulação da vasoconstrição
EPB41	-	-
EPC1	Ligação aos ácidos nucleicos Ligação à cromatina	Transcrição Regulação da transcrição
EPHA2	-	Desenvolvimento do sistema nervoso
ERAP2	Atividade de metalopeptidase	Proteólise
ERBB3	-	Processo apoptótico Proliferação celular Adesão Processo apoptótico Desenvolvimento do sistema nervoso
ERCC3	Atividade de DNA helicase Atividade de hidrolase Ligação aos ácidos nucleicos	Reparação de DNA
ESR1	Atividade de fator de transcrição	Transcrição do promotor da RNA polimerase II Comunicação celular Regulação da transcrição do promotor da RNA polimerase II

Gene	Função Molecular	Processo Biológico
F5	Atividade de oxireductase Atividade de peptidase tipo-serina Atividade de metalopeptidase Atividade de recetor Transportador lipídico Transportador transmembranar Ligação a recetor Atividade de regulador enzimático	Sistema imune Proteólise Transmissão sináptica Adesão Perceção visual e sensorial Desenvolvimento da ectoderme e mesoderme Desenvolvimento do sistema esquelético Angiogénese Desenvolvimento do sistema nervoso Desenvolvimento cardíaco Coagulação sanguínea Transporte lipídico Transporte de proteínas intracelulares Endocitose Transporte de vitaminas Regulação da atividade catalítica
FAM26F	-	-
FAM54A	-	-
FARP1	Atividade catalítica Ligação proteica Atividade regulador GTPase	Processo metabólico Desenvolvimento da mesoderme Desenvolvimento do sistema esquelético Regulação da atividade catalítica
FBXO36	-	-
FBXW5	-	-
FBXW8	-	-
FCHSD2	-	-
FHL2	Fator de transcrição	Desenvolvimento da mesoderme Desenvolvimento muscular
FILIP1	-	-
FKTN	-	-
FLII	Constituinte estrutural do citoesqueleto Ligação à actina	Processo celular Organização dos componentes celulares
FN1	Ligação a recetor	Comunicação celular Adesão
FNDC3A	Atividade de proteína cinase Constituinte estrutural do citoesqueleto Ligação proteica Atividade regulador GTPase	Fosforilação proteica Comunicação celular Adesão Contração muscular Desenvolvimento da mesoderme Desenvolvimento muscular Regulação da atividade catalítica
FOS	Fator de transcrição	Sistema imune Indução de apoptose Transcrição Ciclo celular Sistema neurológico Regulação da transcrição
FOXM1	Atividade peptidase do tipo-cisteína Ligação ao DNA Ligação a recetor Atividade enzimática reguladora	Processo apoptótico Processo catabólico Processo biosintético Processo catabólico de DNA

Gene	Função Molecular	Processo Biológico
FOXRED1	Atividade oxireductase	Cadeia respiratória transportadora de eletrões Metabolismo de aminoácidos Metabolismo proteico e lipídico
FRMD4A	Constituinte estrutural do citoesqueleto	Processo celular Morfogénese Organização dos componentes celulares
FTSJ3	Ligação aos ácidos nucleicos	Metabolismo de rRNA
GAD2	-	Metabolismo dos aminoácidos
GGA3	Atividade de cinase Constituinte estrutural do citoesqueleto Atividade de transportador transmembranar Ligação proteica Regulador de cinase	Processo metabólico Comunicação celular Transporte de lisossomas Endocitose Regulação da atividade catalítica
GMPR2	Atividade oxireductase	Metabolismo das purinas
GNGT2	Atividade GTPase Ligação proteica	Processo metabólico Comunicação celular
GOLGA2	-	-
GOT1	Atividade de transaminase	Metabolismo dos aminoácidos
GPAM	-	-
GPR119	Atividade de recetor acoplado à proteína C	Sinalização celular Perceção visual e sensorial
GPR151	Atividade de recetor acoplado a proteína G	Comunicação celular
GPR17	Atividade de recetor	Processo celular
GPR45	Atividade de recetor acoplado a proteína G	Comunicação celular
GPRC6A	Atividade de recetor de glutamato Canal iónico dependente de ligando	Processo celular Resposta a estímulos Transporte
GRIK3	Atividade de recetor Canal iónico dependente de ligando	Transmissão sináptica Sistema neurológico Transporte de catiões
GRIN3A	Atividade de recetor Canal iónico dependente de ligando	Transmissão sináptica Sistema neurológico
GRN	-	-
GSDMC	-	-
GSTA5	-	-
GTPBP5	Atividade GTPase Ligação proteica	Processo metabólico
GXYLT1	Atividade de transferase	Processo metabólico dos hidratos de carbono Glicosilação proteica
HABP2	Atividade peptidase do tipo-serina	Processo metabólico
HACL1	Atividade de oxireductase Atividade de transferase	Biossíntese de vitaminas Processo metabólico dos hidratos de carbono Biossíntese de aminoácidos
HADH	Atividade oxireductase	Metabolismo de hidratos de carbono β-oxidação dos ácidos gordos

Gene	Função Molecular	Processo Biológico
HDAC1	Atividade de oxireductase Atividade de desacetilase Ligação aos ácidos nucleicos	Processo apoptótico Transcrição Ciclo celular Processo apoptótico Regulação da transcrição Regulação negativa do processo apoptótico Organização da cromatina
HEXA	Atividade de hidrolase	Metabolismo de polissacarídeos Metabolismo lipídico
HIGD1B	-	-
HMCN1	Constituinte estrutural do citoesqueleto	Fosforilação proteica Processo celular Adesão Contração muscular Desenvolvimento da mesoderme Morfogénese
HOXB4	Fator de transcrição	Transcrição Especificação Desenvolvimento da ectoderme e mesoderme do trato digestivo Desenvolvimento do músculo-esquelético Angiogénese Desenvolvimento do sistema nervoso Crescimento muscular
HOXC9	Fator de transcrição	Transcrição Especificação Desenvolvimento da ectoderme e mesoderme Desenvolvimento embrionário Desenvolvimento do sistema esquelético Angiogénese Desenvolvimento do sistema nervoso Desenvolvimento muscular Regulação da transcrição
HPGDS	-	-
HSPA1L	-	Sistema imune *Folding* de proteínas Resposta ao *stress* Biossíntese proteica
HTRA2	Atividade peptidase do tipo-serina	Processo apoptótico *Folding* proteico Proteólise Comunicação celular Processo apoptótico
HYI	-	-
IARS	-	Tradução
IGHMBP2	DNA helicase RNA helicase Atividade endoribonuclease	Replicação de DNA Fator de transcrição mRNA splicing via spliceossoma Processo catabólico do RNA Metabolismo proteico Ciclo celular
INPP5B	-	Metabolismo de fosfolípidos Metabolismo de monossacarídeos Comunicação celular Transporte de proteínas intracelulares

Gene	Função Molecular	Processo Biológico
IQGAP1	Atividade catalítica Ligação proteica Regulação GTPase	Processo metabólico Ciclo celular Comunicação celular Regulação da atividade catalítica
IQGAP3	Atividade catalítica Ligação proteica Atividade de GTPase	Processo metabólico Ciclo celular Comunicação celular Regulação da atividade catalítica
IQSEC1	Atividade de GTPase Atividade de pirofosfatase Regulador da atividade GTPase	Processo metabólico Processo celular Regulação do processo biológico Regulação da atividade catalítica Organização do citoesqueleto
IRAK4	Recetor transmembranar	Sistema imune Fosforilação proteica Comunicação celular
ISX	Fator de transcrição	Transcrição Especificação Desenvolvimento da ectoderme e mesoderme Desenvolvimento do sistema nervoso Desenvolvimento muscular Regulação da transcrição
ITSN1	Atividade catalítica Ligação ao ião cálcio Ligação proteica Atividade de GTPase	Processo metabólico Transmissão sináptica Secreção de neurotransmissores Transporte de proteínas intracelulares Endocitose Regulação da atividade catalítica
KAT2B	Atividade de transferase Ligação aos ácidos nucleicos Ligação à cromatina	Fator de transcrição Processo celular Organização da cromatina
KBTBD8	-	-
KCNJ12	-	-
KCNJ16	-	-
KCNS3	Canal de potássio dependente de voltagem Canal de catiões Atividade de transportador transmembranar	Processo celular Transporte de catiões Regulação de processos biológicos
KIAA0564	-	-
KIAA0895	-	-
KIF23	Atividade motora de microtúbulos Constituinte estrutural do citoesqueleto	Processo metabólico Citocinese Mitose Meiose Segregação de cromossomas Morfogénese Transporte de proteínas intracelulares Transporte mediado por vesículas Organização dos componentes celulares
KLHL31	-	-
KLHL32	-	-

Gene	Função Molecular	Processo Biológico
KRT18	Constituinte estrutural do citoesqueleto	Morfogénese Organização dos componentes celulares
KRT75	Constituinte estrutural do citoesqueleto	Processo celular Morfogénese Organização dos componentes celulares
KYNU	Atividade de hidrolase	Metabolismo de aminoácidos
LAMA1	Atividade de recetor	Comunicação celular Adesão Sistema neurológico Desenvolvimento da ectoderme Desenvolvimento do sistema nervoso
LAMA4	Atividade de recetor	Comunicação celular Adesão Sistema neurológico Desenvolvimento da ectoderme Desenvolvimento do sistema nervoso
LARP4	Ligação ao RNA	Processo metabólico
LASS5	-	-
LENG1	Atividade de recetor	Processo celular
LIMK1	Regulador da Atividade tirosina-cinase	Formação de gâmetas femininos Sistema imune Indução da apoptose Processo metabólico Fosforilação proteica Ciclo celular Comunicação celular Morfogénese Desenvolvimento embrionário Resposta ao *stress* Organização dos componentes celulares
LRP2	Atividade catalítica Ligação proteica Atividade de regulador enzimático	Processo metabólico Comunicação celular Regulação da atividade catalítica Resposta a estímulos Regulação do processo biológico
LRRC31	-	-
LRRC8E	Atividade de cinase Atividade de recetor Fator de crescimento	Sistema imune Comunicação celular Regulação da atividade catalítica
LRRK2	Atividade de tirosina-cinase	Processo metabólico Fosforilação proteica Ciclo celular Comunicação celular Resposta ao stress Organização dos componentes celulares
LRRN4	Atividade de recetor	Processo do sistema imune Comunicação celular
LYST	Ligação proteica	Comunicação celular
MAN2C1	-	-
MAP1S	Constituinte estrutural do citoesqueleto Ligação aos microtúbulos	Processo celular

Gene	Função Molecular	Processo Biológico
MAP2K3	-	-
MBTPS1	Atividade peptidase do tipo-serina	Metabolismo proteico
MED27	Atividade de cofator na transcrição Fator de transcrição de ligação do DNA Ligação proteica	Transcrição do promotor da RNA polimerase II Regulação da transcrição do promotor da RNA polimerase II
MIER1	Atividade de fator de transcrição Ligação à cromatina	Transcrição do promotor da RNA polimerase II Processo celular Desenvolvimento da mesoderme
MIPEP	Atividade de metalopeptidase	Proteólise Processo celular Transporte mitocondrial Transporte proteico intracelular Organização da mitocôndria
MLL5	-	-
MMADHC	-	-
MORN3	Atividade de cinase	Processo metabólico e celular
MPND	Co-fator de transcrição Ligação à cromatina	Transcrição
MPP3	Atividade de cinase	Processo metabólico Localização
MRP63	-	-
MRPL28	Constituinte estrutural do ribossoma Ligação aos ácidos nucleicos	Tradução
MTCH2	-	Transporte
MTMR12	-	Espermatogénese Processo metabólico fosfolipídico Transporte proteico intracelular Transporte mediado por vesículas
MTTP	Transporte de lípidos Transporte transmembranar	Processo metabólico lipídico Transporte de lípidos
MYO3A	Atividade motora Constituinte estrutural do citoesqueleto Ligação proteica Atividade enzimática	Processo metabólico Citocinese Movimento dos componentes celulares Mitose Comunicação celular Contração muscular Perceção sensorial Desenvolvimento da mesoderme Morfogénese Crescimento muscular Transporte de proteínas intracelulares Transporte mediado por vesículas Regulação da atividade catalítica Organização dos componentes celulares
MYOF	-	Contração muscular
NAALADL1	-	-
NADK	Atividade de cinase	Processo catabólico
NCSTN	-	-

Gene	Função Molecular	Processo Biológico
NDOR1	Atividade de oxireductase	Processo metabólico
NHS	-	-
NOTCH1	Atividade de recetor Ligação a recetor	Sistema imune Transcrição Sinalização celular Desenvolvimento da ectoderme Desenvolvimento do sistema nervoso Regulação da transcrição
NPNT	Constituinte estrutural de matriz extracelular	-
NT5E	Atividade de hidrolase Atividade de fosfatase	Processo metabólico
NUDT13	-	-
NUP155	-	-
OR10C1	-	-
OR1L4	-	-
OR1L6	-	-
OR1N2	-	-
OR52J3	-	-
OR5I1	-	-
OR5P2	-	-
OR5P2	-	-
OR8B4	-	-
OTOF	-	Contração muscular
PABPC1	Atividade de fator de transcrição Atividade catalítica Replicação de DNA Ligação a RNA (poli-A)	Replicação de DNA Transcrição mRNA splicing via spliceossoma Poliadenilação e mRNA Processo metabólico proteico Ciclo celular Desenvolvimento da ectoderme Desenvolvimento do sistema nervoso
PADI3	Atividade de hidrolase	Modificação proteica
PAM	Atividade oxireductase	Modificação proteica
PAX7	Fator de transcrição	Transcrição Especificação Desenvolvimento da ectoderme e mesoderme Desenvolvimento do sistema esquelético Desenvolvimento do sistema nervoso Desenvolvimento muscular Regulação da transcrição
PBRM1	-	-
PCP2	Atividade catalítica Ligação proteica Atividade regulador GTPase	Processo metabólico Citocinese Mitose Sinalização celular Regulação da atividade catalítica
PDCD11	-	-

Gene	Função Molecular	Processo Biológico
PDE1A	Atividade de hidrolase	Processo metabólico Comunicação celular Perceção sensorial e visual
PDE4C	Atividade de hidrolase	Processo metabólico Comunicação celular Perceção sensorial e visual
PDE6A	Atividade de hidrolase	Processo metabólico Comunicação celular Perceção sensorial e visual
PDZD2	Ligação a recetores	Resposta imune Perceção visual Desenvolvimento do sistema nervoso Resposta a estímulos
PGM1	Atividade de transferase intramolecular	Geração de percursores de metabolitos e energia Processo catabólico Processo metabólico do glicogénio
PHLPP2	Atividade de cinase Atividade de recetor Fator de crescimento Regulador da atividade de cinase	Sistema imune Processo metabólico Comunicação celular Regulação da atividade catalítica
PIDD	Atividade de cinase Atividade de recetor Fator de crescimento Regulador da atividade de cinase	Sistema imune Processo metabólico Comunicação celular Regulação da atividade catalítica
PIGT	-	-
PKD1L2	Canal iónico Transportador transmembranar	Processo celular Resposta a estímulos externos Resposta a estímulos abióticos Transporte de catiões
PKD1L2	Atividade de canal iónico Transporte transmembranar de catiões	Processo celular Resposta a estímulos externos Resposta a estímulos abióticos Transporte de catiões
PKD2L1	Canal iónico Atividade de transportador transmembranar	Processo celular Resposta a estímulos externos Resposta a estímulos abióticos Transporte de catiões
PLB1	-	-
PLBD1	-	-
PLEKHG2	Atividade catalítica Ligação a recetor Regulação GTPase	Imunidade mediada por células B Processo metabólico Comunicação celular Sistema neurológico Defesa celular Regulação da atividade catalítica
PLEKHH3	Atividade motora Constituinte estrutural do citoesqueleto Ligação proteica Regulador enzimático	Processo metabólico Citocinese Movimento dos componentes celulares Mitose Comunicação celular Contração muscular Perceção sensorial Desenvolvimento da mesoderme Morfogénese

Gene	Função Molecular	Processo Biológico
PLXDC1	-	-
PLXNA2	Recetor transmembranar Atividade de tirosina-cinase	Fosforilação proteica Comunicação celular Desenvolvimento do sistema nervoso
PNCK	Atividade cinase Ligação ao ião cálcio Ligação à calmodulina	Fosforilação proteica Comunicação celular
POLQ	Atividade de DNA helicase Atividade de hidrolase Atividade de peptidase Ligação a mRNA	RNA *splicing* Ciclo celular Proteólise
POLR3E	Atividade de RNA polimerase Ligação aos ácidos nucleicos	Transcrição
POMT2	Atividade de transferase	Tradução Glicosilação proteica
PPFIBP2	-	-
PPYR1	Atividade de recetor acoplado a proteína G	Sistema imune Comunicação celular Contração muscular Circulação sanguínea Perceção sensorial Resposta ao *stress*
PRF1	Atividade peptidase do tipo serina Metalopeptidase Atividade de recetor Transportador lipídico	Ativação do complemento Proteólise Comunicação celular Adesão Coagulação sanguínea Transporte de lípidos
PRKD1	-	-
PRR16	-	-
PSMB10	Atividade de fosforilase	Metabolismo do glicogénio
PSMC2	Atividade de hidrolase	Proteólise
PSPH	Atividade de fosfatase Ligação	Processo metabólico Biossíntese de aminoácidos Processo celular
PTPN21	Atividade de fosfatase	Modificação proteica
PYGL	-	Reparação do DNA Ciclo celular
RAB3IP	Atividade catalítica Ligação proteica Atividade reguladora de GTPase	Processo metabólico Transmissão sináptica Sistema neurológico Regulação da atividade catalítica
RAD17	Atividade de cinase	Processo metabólico rRNA Fosforilação proteica
RELL2	-	-
RGL1	Atividade catalítica Ligação proteica Regulação GTPase	Processo metabólico Mitose Comunicação celular Regulação da atividade catalítica
RIOK2	Atividade ligase	Proteólise
RNF157	Atividade de endoribonuclease Ligação aos ácidos nucleicos	Processo metabólico tRNA Tradução

Gene	Função Molecular	Processo Biológico
RPP30	Atividade de endoribonuclease Ligação aos ácidos nucleicos	Processo metabólico tRNA
RPP38	Constituinte estrutural do ribossoma Ligação aos ácidos nucleicos	Processo metabólico
RPS20	Constituinte estrutural do ribossoma Ligação aos ácidos nucleicos	Tradução
RRS1	-	-
RSPH4A	Constituinte estrutural do citoesqueleto	Movimento dos componentes celulares
SCUBE2	-	Processo celular Desenvolvimento do sistema esquelético
SCYL2	Transportador transmembranar	Metabolismo do colesterol Transporte de proteínas intracelulares
SEC14L3	Atividade de hidrolase Ligação proteica Atividade de regulador GTPase	Processo metabólico Processo celular Morfogénese Transporte de proteínas intracelulares Exocitose Regulação da atividade catalítica Organização dos componentes celulares
SEL1L	Atividade catalítica Ligação proteica Atividade enzimática	Processo metabólico Regulação da atividade catalítica
SELENBP1	-	-
SEPT4	Atividade GTPase Constituinte estrutural do citoesqueleto Ligação proteica	Processo metabólico Citocinese Mitose
SERPINB8	Atividade de peptidase tipo-serina Atividade inibidora de peptidase	Proteólise Regulação do processo biológico Regulação da atividade catalítica
SEZ6L2	Atividade de peptidase tipo-serina Atividade de metalopeptidase Atividade de recetor	Ativação do complemento Proteólise Comunicação celular Adesão Coagulação sanguínea Transporte de lípidos
SF3B3	Atividade catalítica Ligação a RNA (Poli-A)	Reparação de DNA Poliadenilação de mRNA RNA *splicing*
SGSM3	Atividade de hidrolase Ligação proteica Atividade reguladora de GTPase	Processo metabólico e celular Morfogénese Transporte de proteínas intracelulares Exocitose Regulação da atividade catalítica Organização dos componentes celulares
SH2B3	-	Imunidade mediada por células B Comunicação celular Resposta de defesa celular
SHC1	Ligação a recetor	Processo apoptótico Comunicação celular Desenvolvimento da ectoderme

Gene	Função Molecular	Processo Biológico
SIGLEC1	Atividade de recetor Constituinte estrutural da barreira de mielina	Imunidade mediada por células B Desenvolvimento do sistema nervoso Resposta a estímulos
SLC11A1	Atividade de transportador transmembranar	Sistema imune Resposta ao *stress* Transporte de catiões
SLC34A3	Transporte transmembranar Ligação proteica	Processo metabólico Transporte iões
SLC45A2	Atividade de transportador transmembranar	Metabolismo de hidratos de carbono Transporte de hidratos de carbono
SLC4A3	Atividade de transportador transmembranar	Morfogénese Transporte iónico Organização dos componentes celulares
SLC5A12	Transportador transmembranar Transportador iónico	Metabolismo de hidratos de carbono Metabolismo de aminoácidos Transporte de catiões Transporte extracelular Transporte de aminoácidos
SLC9A3R1	Atividade de recetor	Sistema imune Transcrição Fosforilação proteica Movimento dos componentes celulares Sinalização celular Adesão Perceção sensorial Desenvolvimento da ectoderme Desenvolvimento do sistema nervoso Regulação da transcrição
SLIT3	Ligação aos ácidos nucleicos	Reparação do DNA
SMARCAL1	Atividade de DNA helicase Ligação aos ácidos nucleicos	Reparação de DNA Recombinação de DNA Transcrição Processo celular Regulação da transcrição Organização da cromatina
SMC6	-	Transcrição
SNAPC1	Atividade de recetor Transportador transmembranar Ligação a lípidos	Processo metabólico lipídico Processo celular Transporte lipídico Transporte de proteínas intracelulares Endocitose
SORL1	Atividade de transferase	Metabolismo de polissacarídeo Processo metabólico Glicosilação proteica
SREBF1	Fator de transcrição	Transcrição Processo metabólico lipídico
SRPRB	-	-
ST6GAL2	-	-
STAB2	Constituinte estrutural da matriz extracelular	Comunicação celular Adesão
STEAP4	-	Transporte de proteínas intracelulares Pinocitose
STON2	Atividade de transferase	Processo metabólico

Gene	Função Molecular	Processo Biológico
STX2	Ligação proteica	Processo celular Transporte proteico Transporte mediado por vesículas Localização proteica
STXBP2	Transporte transmembranar	Transmissão sináptica Secreção de neurotransmissores Transporte lisossomal Transporte de proteínas intracelulares Exocitose da vesícula sináptica
SULT2A1	Constituinte estrutural do citoesqueleto Ligação á actina	Movimento dos componentes celulares Morfogénese Organização dos componentes celulares
SVIL	Constituinte estrutural do citoesqueleto Ligação ao ião cálcio Ligação à actina	Processo celular Morfogénese Organização dos componentes celulares
SYNE2	Constituinte estrutural do citoesqueleto Ligação à actina	Movimento dos componentes celulares Morfogénese Organização dos componentes celulares
TACR2	Atividade de recetor acoplado a proteína G	Sistema imune Comunicação celular Contração muscular Perceção sensorial Resposta ao *stress*
TBCK	Atividade de hidrolase	Processo metabólico e celular Transporte de proteínas intracelulares Exocitose Regulação da atividade catalítica Organização dos componentes celulares
TBL3	Atividade de recetor	Processo celular
TBX15	Fator de transcrição	Transcrição Desenvolvimento da mesoderme Regulação da transcrição
TCN2	Ligação à cobalamina	Processo metabólico da cobalamina Metabolismo da cobalamina
TEP1	Atividade de cinase Atividade de hidrolase Ligação a mRNA Ligação proteica Atividade inibidora de cinase	RNA *splicing* Localização de RNA Regulação da atividade catalítica
TGIF1	Atividade de RNA polimerase dirigida por DNA	Transcrição do promotor da RNA polimerase II
TIAM2	Atividade catalítica Ligação a recetores	Imunidade mediada por células B Processo metabólico Comunicação celular Sistema neurológico Defesa celular Regulação da atividade catalítica
TIMM44	Canal iónico	Transporte
TMC7	Canal iónico	Transporte
TMEM144	-	-
TNFAIP3	Atividade de hidrolase Ligação ao DNA de cadeia dupla	Proteólise

Gene	Função Molecular	Processo Biológico
TP73	Ligação aos ácidos nucleicos Fator de transcrição Atividade cinase Ligação ao DNA Ligação proteica	Processo apoptótico Processo biosintético Transcrição Fosforilação proteica Ciclo celular Comunicação celular Resposta ao *stress* Resposta a estímulos abióticos Regulação da transcrição Processo metabólico Regulação negativa do processo apoptótico Regulação do ciclo celular Regulação da atividade catalítica
TPMT	Atividade de transferase	Metilação Resposta à testosterona Processo metabólico de xenobiótico
TRAF3	Ligação a recetor	Sistema imune Indução da apoptose Comunicação celular Desenvolvimento da mesoderme Indução da apoptose Desenvolvimento do sistema esquelético
TRAP1	-	*Folding* de proteínas Resposta ao *stress*
TRERF1	Ligação ao DNA	Processo metabólico Processo catabólico
TRIM41	Atividade de ligase	Processo de modificação de proteínas celulares
TRMT11	Atividade de transferase Ligação ao tRNA	Processamento de tRNA
TSEN54	-	Processamento de mRNAs
TSR1	-	Processo metabólico
TTBK2	Atividade de cinase	Reparação de DNA Fosforilação proteica Citocinese Mitose Meiose Segregação de cromossomas Comunicação celular Morfogénese Transporte proteico intracelular Endocitose Organização dos componentes celulares
TTC15	Atividade de isomerase	*Folding* das proteínas Modificação proteica
TTLL12	Atividade de ligase Constituinte estrutural do citoesqueleto	Processo metabólico proteico
TUBB1	Constituinte estrutural do citoesqueleto	Movimento dos componentes celulares Mitose Segregação de cromossomas Morfogénese Transporte de proteínas intracelulares Organização dos componentes celulares
TXNDC11	-	Sistema imune

Gene	Função Molecular	Processo Biológico
UNC45A	-	Sistema imune *Folding* das proteínas Resposta ao *stress*
UNC45B	-	-
UPF3A	-	Processo catabólico
UQCRH	Atividade oxireductase	Fosforilação oxidativa Cadeia respiratória transportadora de eletrões
USH2A	Atividade de recetor	Comunicação celular Adesão Sistema neurológico Desenvolvimento da ectoderme Desenvolvimento do sistema nervoso
UTP20	Fator de transcrição	Transcrição Processo celular Desenvolvimento da mesoderme Desenvolvimento do sistema esquelético Regulação da transcrição
VDR	Fator de transcrição	Transcrição Processo celular Desenvolvimento da mesoderme Desenvolvimento do sistema esquelético Regulação da transcrição
WDHD1	Ligação ao DNA	Transcrição do promotor da RNA polimerase II Regulação da transcrição
WDR31	-	-
WDR33	Ligação a mRNA	Reparação de DNA Processamento de mRNA
WDSUB1	-	-
XRCC1	-	-
YWHAB	-	Ciclo celular Comunicação Celular
ZFAND2B	Ligação ao RNA	-
ZFYVE27	-	-
ZKSCAN2	Fator de transcrição	Processo metabólico Processo biosintético Transcrição Processo celular Regulação da transcrição
ZNF7	-	-

Anexo 12: Distribuição do número de rSNPs identificados nos cromossomas dos genomas H1, H2, H3 e H4.

H1	SNPs intrónicos	rSNPs na região intrónica	LD com rSNPs	Regulação Proximal	Regulação Distal	Regulação mediada por uma proteína de ligação ao RNA
1	**100.108**	79.993	72.840	16.780	29.950	71.646
2	**90.770**	72.580	69.250	13.304	21.497	65.490
3	**83.542**	65.489	61.392	11.339	17.850	60.185
4	**66.883**	52.953	49.905	8.263	11.214	50.022
5	**58.672**	48.849	45.216	9.002	12.885	44.918
6	**70.691**	23.528	19.501	3.110	6.094	22.115
7	**73.424**	56.612	51.066	9.596	16.129	51.534
8	**60.100**	46.864	44.843	7.774	11.190	43.258
9	**48.707**	40.611	36.609	6.631	12.374	37.450
10	**65.982**	55.198	49.328	8.396	14.051	51.503
11	**61.684**	48.346	45.260	11.932	16.206	43.379
12	**56.735**	46.938	43.400	10.713	14.699	43.186
13	**33.802**	28.302	25.459	3.681	5.438	27.355
14	**32.872**	27.819	25.476	6.693	10.732	25.149
15	**38.856**	29.274	28.677	7.251	10.695	25.939
16	**43.881**	36.900	31.348	9.216	14.680	33.380
17	**45.276**	36.672	30.684	11.894	16.971	32.039
18	**28.161**	23.898	21.398	4.771	5.498	22.478
19	**32.720**	25.466	21.611	10.424	15.945	20.747
20	**28.372**	23.217	21.032	4.179	8.711	21.429
21	**16.828**	12.202	10.324	3.075	5.655	10.553
22	**19.937**	15.807	14.362	4.430	7.262	13.789
X	**15.255**	11.455	52.772	1.535	1.989	10958
Y	**145**	35	0	1	0	34
Total:	**1.173.403**	909.008	871.753	183.990	287.715	828536

H2	SNPs intrónicos	rSNPs na região intrónica	LD com rSNPs	Regulação Proximal	Regulação Distal	Regulação mediada por uma proteína de ligação ao RNA
1	**115.923**	95.653	85.433	19.936	34.347	85.313
2	**110.052**	88.276	82.552	16.048	25.168	79.756
3	**96.082**	75.347	69.928	13.429	20.338	69.455
4	**79.293**	63.285	59.112	9.710	13.106	60.064
5	**66.543**	55.960	51.603	10.613	14.481	51.276
6	**81.579**	67.747	55.842	9.277	17.277	63.232
7	**85.072**	64.660	59.495	10.674	17.706	59.015
8	**69.350**	54.564	51.788	8.781	12.293	50.389
9	**55.079**	45.652	40.781	7.399	13.453	42.243
10	**74.724**	61.882	55.080	9.013	15.184	57.930
11	**69.006**	53.762	50.418	13.062	17.214	48.451
12	**64.625**	53.394	49.143	12.012	16.264	49.134
13	**38.057**	31.782	28.931	4.020	5.966	30.811
14	**37.391**	31.252	28.938	7.302	11.457	28.437
15	**45.100**	33.594	32.627	8.441	12.397	29.588
16	**48.297**	40.614	34.432	10.843	16.600	36.497
17	**50.481**	41.064	34.203	13.143	19.398	35.510
18	**31.505**	26.142	23.496	51.65	6.011	24.597
19	**35.475**	26.979	22.817	11.239	16.747	22.043
20	**29.029**	23.568	21.334	4.159	8.620	21.599
21	**18.137**	12.718	10.839	3.209	5.972	10.884
22	**20.801**	16.123	14.648	4.668	7.604	13.871
X	**18.222**	13.343	5.929	1.771	1.981	12.852
Y	**316**	75	0	6	0	70
Total:	**1.340.139**	1.077.436	969.369	213.920	329.584	983.017

H3	SNPs intrónicos	rSNPs na região intrónica	LD com rSNPs	Regulação Proximal	Regulação Distal	Regulação mediada por uma proteína de ligação ao RNA
1	**112.715**	93.884	83.089	19.491	33.922	83.617
2	**105.228**	84.907	79.942	15.760	24.731	76.703
3	**97.706**	76.347	71.314	12.941	19.650	70.338
4	**78.666**	63.740	58.987	10.003	13.572	60.158
5	**64.606**	54.962	50.636	9.765	13.943	50.902
6	**85.399**	65.963	55.850	9.097	16.453	61.556
7	**82.511**	62.621	57.340	10.158	16.975	56.883
8	**70.556**	55.233	52.715	8.888	12.545	51.060
9	**54.479**	45.215	40.774	7.432	12.931	41.886
10	**72.705**	60.922	53.805	8.736	14.744	56.991
11	**69.100**	53.435	50.251	12.954	17.033	48.030
12	**62.841**	52.433	47.340	11.829	15.628	48.305
13	**38.858**	32.229	29.285	4.250	6.108	31.199
14	**37.730**	31.555	28.820	7.456	11.630	28.594
15	**44.782**	33.990	33.158	8.536	12.588	30.237
16	**48.478**	40.810	34.454	10.458	15.894	36.855
17	**47.660**	38.523	32.129	12.421	18.249	33.519
18	**32.648**	27.074	24.707	5.151	6.005	25.513
19	**34.206**	26.692	22.607	11.196	16.844	21.856
20	**28.694**	23.200	20.930	3.941	8.139	21.512
21	**17.772**	12.666	10.718	3.201	5.758	10.999
22	**21.184**	16.574	15.320	4.889	7.655	14.259
X	**18.448**	13.537	6.282	1.831	2.016	12.989
Y	**118**	25	0	2	0	23
Total:	**1.327.090**	1.066.537	960.453	210.386	323.013	973.984

H4	SNPs intrónicos	rSNPs na região intrónica	LD com rSNPs	Regulação Proximal	Regulação Distal	Regulação mediada por uma proteína de ligação ao RNA
1	**104.236**	86.532	77.058	18.411	31.312	77.235
2	**95.933**	75.854	72.091	13.920	22.217	68.735
3	**88.658**	69.500	64.892	12.062	18.480	64.183
4	**68.816**	55.393	51.651	8.882	12.457	52.333
5	**61.097**	51.223	46.871	9.116	13.010	47.299
6	**74.098**	55.756	47.855	7.542	13.951	51.949
7	**74.667**	56.565	51.585	9.986	16.618	51.048
8	**61.193**	47.443	45.807	7.979	11.248	43.385
9	**53.755**	44.532	40.248	7.466	13.373	40.809
10	**69.116**	57.644	51.035	8.630	14.181	53.864
11	**63.449**	48.987	46.233	11.760	15.980	44.006
12	**58.804**	48.178	43.694	11.011	15.032	44.336
13	**36.132**	29.773	26.693	3.814	5.580	28.824
14	**35.987**	29.902	27.859	7.026	11.207	27.067
15	**41.315**	31.230	30.436	7.711	11.501	27.925
16	**47.002**	39.651	33.782	10.426	15.891	35.901
17	**46.542**	37.176	31.200	11.993	17.732	32.317
18	**30.150**	25.118	22.612	5.028	5.644	23.685
19	**34.686**	26.491	22.155	11.164	16.601	21.745
20	**27.828**	22.523	20.092	3.914	8.173	20.867
21	**16.616**	11.958	10.191	3.077	5.469	10.414
22	**20.563**	16.003	14.316	4.850	7.666	13.794
X	**16.563**	12.240	5.425	1.669	2.099	11.819
Y	**102**	25	0	1	0	24
Total	**1.227.308**	979.697	883.781	197.438	305.422	893.564

Anexo 13: Descrição dos miRNAs, nos quais foram identificados SNPs, nos genomas H1, H2, H3 e H4.

H1					
Cromossoma	miRNA	Localização	Ref	Alt	Zigotia
1	hsa-mir-3118-1	142667306	C	T	het
1	hsa-mir-3118-2	143163799	G	A	het
1	hsa-mir-3118-2	143163816	G	A	het
2	hsa-mir-559	47604856	T	C	het
2	hsa-mir-1302-3	114340663	G	A	het
2	hsa-mir-663b	133014552	C	T	het
2	hsa-mir-663b	133014556	C	T	het
2	hsa-mir-663b	133014580	C	T	het
2	hsa-mir-663b	133014591	C	T	het
2	hsa-mir-663b	133014592	G	A	het
2	hsa-mir-663b	133014602	G	C	het
2	hsa-mir-663b	133014603	G	T	het
2	hsa-mir-663b	133014612	A	C	het
2	hsa-mir-663b	133014617	A	G	het
2	hsa-mir-663b	133014619	C	T	het
2	hsa-mir-663b	133014621	A	C	het
2	hsa-mir-663b	133014633	T	C	het
2	hsa-mir-663b	133014640	C	T	het
2	hsa-mir-933	176032376	G	A	het
2	hsa-mir-4268	220771223	C	T	het
3	hsa-mir-3135	20179097	C	T	het
3	hsa-mir-1324	75679930	A	C	het
3	hsa-mir-1324	75679937	G	A	het
3	hsa-mir-1324	75679938	T	C	het
3	hsa-mir-1324	75679944	C	T	het
3	hsa-mir-1324	75679945	C	T	het
3	hsa-mir-1324	75679958	C	T	het
3	hsa-mir-1324	75679965	C	A	het
3	hsa-mir-1324	75679973	A	G	het
3	hsa-mir-1324	75679975	C	T	het
3	hsa-mir-1324	75679988	A	G	het
3	hsa-mir-1324	75679990	G	A	het
3	hsa-mir-1324	75679997	C	T	het
3	hsa-mir-4273	75787464	C	T	het
3	hsa-mir-4273	75787476	C	T	het
3	hsa-mir-4273	75787486	C	T	het
3	hsa-mir-4273	75787511	C	A	het
4	hsa-mir-943	1988193	T	C	het
4	hsa-mir-1255b-1	36428048	G	A	het
4	hsa-mir-1255a	102251501	T	C	het
5	hsa-mir-4277	1708983	G	A	hom
5	hsa-mir-1274a	41475766	C	T	hom
5	hsa-mir-449c	54468124	A	T	het
5	hsa-mir-3141	153975576	G	A	het
5	hsa-mir-146a	159912418	C	G	het
5	hsa-mir-585	168690612	A	G	het
6	hsa-mir-548u	57254939	G	T	het
6	hsa-mir-548u	57254955	G	A	het
6	hsa-mir-548u	57254986	G	A	het
6	hsa-mir-548u	57255001	C	A	het

8	hsa-mir-1206	129021179	G	A	hom
9	hsa-mir-1299	69002248	C	T	het
9	hsa-mir-1299	69002251	A	G	het
9	hsa-mir-1299	69002271	T	A	het
9	hsa-mir-1299	69002274	G	T	het
9	hsa-mir-1299	69002283	A	T	het
9	hsa-mir-1299	69002294	C	T	het
9	hsa-mir-1299	69002295	G	A	het
9	hsa-mir-1299	69002304	C	G	het
9	hsa-mir-1299	69002317	T	C	het
10	hsa-mir-605	53059406	T	C	het
10	hsa-mir-608	102734778	C	G	het
10	hsa-mir-2110	115933905	C	T	het
10	hsa-mir-202	135061112	C	T	hom
11	hsa-mir-1304	93466866	G	T	hom
11	hsa-mir-3167	126858392	A	G	het
11	hsa-mir-3167	126858406	A	T	het
12	hsa-mir-4302	26026988	G	A	het
12	hsa-mir-618	81329536	A	C	hom
14	mgU6-53B	21860360	G	A	hom
14	hsa-mir-3172	32954336	T	C	hom
14	hsa-mir-412	101531854	A	G	het
15	hsa-mir-3175	93447631	T	G	hom
15	hsa-mir-1302-10	102500741	C	A	hom
15	hsa-mir-1302-10	102500754	C	T	hom
15	hsa-mir-1302-10	102500789	G	A	hom
16	hsa-mir-3176	593277	T	G	het
16	hsa-mir-1826	33965533	A	C	het
16	hsa-mir-1826	33965537	C	T	het
16	hsa-mir-1826	33965538	A	G	het
16	hsa-mir-1826	33965542	C	T	het
16	hsa-mir-1826	33965554	A	G	het
16	hsa-mir-1826	33965560	T	A	het
16	hsa-mir-1826	33965561	T	C	het
16	hsa-mir-1826	33965570	G	A	het
16	hsa-mir-1826	33965582	C	T	het
16	hsa-mir-1826	33965589	G	A	hom
16	hsa-mir-1826	33965591	A	C	het
17	hsa-mir-3183	925764	A	G	het
17	hsa-mir-548h-3	13446924	G	A	hom
17	hsa-mir-2117	41522213	T	A	het
19	hsa-mir-3188	18392894	C	T	het
19	hsa-mir-3188	18392913	A	G	het
19	hsa-mir-320e	47212593	A	G	het
19	hsa-mir-518a-2	54242655	G	A	het
20	hsa-mir-663	26188824	T	C	het
20	hsa-mir-663	26188912	A	C	het
20	hsa-mir-3196	61870141	G	A	het
20	hsa-mir-3196	61870167	C	A	het
20	hsa-mir-4326	61918164	C	G	het
21	hsa-mir-3156-3	14778721	A	T	hom
22	hsa-mir-650	23165340	C	G	het
X	hsa-mir-532	49767832	A	G	hom
X	hsa-mir-532	49767835	A	G	hom

H2					
Cromossoma	miRNA	Localização	Ref	Alt	Zigotia
1	hsa-mir-3117	67094171	G	A	het
1	hsa-mir-3118-2	143163772	G	C	het
1	hsa-mir-3118-2	143163775	A	G	het
1	hsa-mir-3118-2	143163790	T	C	het
1	hsa-mir-3118-2	143163799	G	A	het
1	hsa-mir-3118-2	143163816	G	A	het
2	hsa-mir-1302-3	114340628	C	T	het
2	hsa-mir-1302-3	114340663	G	A	het
2	hsa-mir-663b	133014552	C	T	het
2	hsa-mir-663b	133014556	C	T	het
2	hsa-mir-663b	133014580	C	T	het
2	hsa-mir-663b	133014592	G	A	het
2	hsa-mir-663b	133014602	G	C	het
2	hsa-mir-663b	133014612	A	C	het
2	hsa-mir-663b	133014617	A	G	het
2	hsa-mir-663b	133014619	C	T	het
2	hsa-mir-663b	133014640	C	T	het
2	hsa-mir-3130-3,hsa-mir-3130-4,hsa-mir-3130-1,hsa-mir-3130-2	207647981	C	T	het
2	hsa-mir-4268	220771223	C	T	hom
2	hsa-mir-149	241395503	T	C	het
3	hsa-mir-3135	20179097	C	T	het
3	hsa-mir-1324	75679914	C	T	het
3	hsa-mir-1324	75679915	C	T	het
3	hsa-mir-1324	75679930	A	C	het
3	hsa-mir-1324	75679938	T	C	het
3	hsa-mir-1324	75679945	C	T	het
3	hsa-mir-1324	75679963	C	T	het
3	hsa-mir-1324	75679965	C	A	het
3	hsa-mir-1324	75679973	A	G	het
3	hsa-mir-1324	75679975	C	T	het
3	hsa-mir-1324	75679988	A	G	het
3	hsa-mir-1324	75679990	G	A	het
3	hsa-mir-1324	75679997	C	T	het
3	hsa-mir-4273	75787455	A	G	het
3	hsa-mir-4273	75787462	G	A	het
3	hsa-mir-4273	75787463	C	T	het
3	hsa-mir-4273	75787464	C	T	het
3	hsa-mir-4273	75787474	G	A	het
3	hsa-mir-4273	75787476	C	A	het
3	hsa-mir-4273	75787486	C	T	het
3	hsa-mir-4273	75787490	A	T	het
3	hsa-mir-4273	75787511	C	A	het
3	hsa-mir-944	189547735	T	C	het
4	hsa-mir-943	1988193	T	C	hom
4	hsa-mir-1269	67142620	G	A	hom
5	hsa-mir-4277	1708983	G	A	hom
5	hsa-mir-1274a	41475766	C	T	het
5	hsa-mir-146a	159912418	C	G	het
5	hsa-mir-585	168690612	A	G	het
6	hsa-mir-548a-1	18572056	T	G	het
6	hsa-mir-548u	57254955	G	A	het
6	hsa-mir-548u	57254986	G	A	het

6	hsa-mir-548u	57255001	C	A	het
6	hsa-mir-3144	120336327	C	A	het
6	hsa-mir-3144	120336384	G	A	het
8	hsa-mir-1206	129021179	G	A	het
8	hsa-mir-1234	145625538	C	G	hom
9	hsa-mir-1299	69002271	T	A	het
9	hsa-mir-1299	69002280	A	T	het
9	hsa-mir-1299	69002283	A	T	het
9	hsa-mir-1299	69002284	T	C	het
9	hsa-mir-1299	69002294	C	T	het
9	hsa-mir-1299	69002295	G	A	het
9	hsa-mir-1299	69002304	C	G	het
9	hsa-mir-1299	69002317	T	C	het
10	hsa-mir-604	29833998	A	G	het
10	hsa-mir-604	29834003	G	A	het
10	hsa-mir-938	29891260	C	T	het
10	hsa-mir-608	102734778	C	G	het
10	hsa-mir-1307	105154089	A	G	het
10	hsa-mir-202	135061112	C	G	het
10	hsa-mir-202	135061112	C	T	het
11	hsa-mir-1908	61582708	T	C	het
11	hsa-mir-612	65211979	G	A	het
11	hsa-mir-1304	93466866	G	T	hom
12	hsa-mir-196a-2	54385599	C	T	hom
12	hsa-mir-618	81329536	A	C	hom
13	hsa-mir-4305	40238175	C	T	het
14	hsa-mir-300	101507727	C	T	hom
14	hsa-mir-323b	101522556	T	C	het
15	hsa-mir-3118-4	21038129	T	A	het
15	hsa-mir-3118-4	21038131	C	T	het
15	hsa-mir-3118-4	21038135	A	T	het
15	hsa-mir-3175	93447631	T	G	het
16	hsa-mir-1826	33965533	A	C	het
16	hsa-mir-1826	33965537	C	T	het
16	hsa-mir-1826	33965538	A	G	het
16	hsa-mir-1826	33965542	C	T	het
16	hsa-mir-1826	33965554	A	G	het
16	hsa-mir-1826	33965560	T	A	het
16	hsa-mir-1826	33965561	T	C	het
16	hsa-mir-1826	33965570	G	A	het
16	hsa-mir-1826	33965582	C	T	het
16	hsa-mir-1826	33965589	G	A	het
16	hsa-mir-1826	33965591	A	C	het
16	hsa-mir-1972-2	70064261	C	T	het
17	hsa-mir-548h-3	13446924	G	A	hom
17	hsa-mir-423	28444183	A	C	hom
17	hsa-mir-2117	41522213	T	A	hom
19	hsa-mir-27a	13947292	T	C	hom
19	hsa-mir-3188	18392966	A	G	het
19	hsa-mir-320e	47212593	A	G	het
20	hsa-mir-663	26188824	T	C	hom
20	hsa-mir-663	26188912	A	C	hom
20	hsa-mir-3196	61870167	C	A	het
20	hsa-mir-4326	61918164	C	G	hom
20	hsa-mir-941-1	62550824	G	A	het
20	hsa-mir-941-1	62550880	G	A	het

20	hsa-mir-941-3	62551281	C	G	het
20	hsa-mir-647	62574006	A	G	hom
21	hsa-mir-3156-3	14778721	A	T	hom
21	hsa-mir-3118-5	15017156	T	A	het

22	hsa-mir-650	23165340	C	G	het
22	hsa-mir-658	38240368	G	C	het
X	hsa-mir-532	49767832	A	G	hom
X	hsa-mir-532	49767835	A	G	hom

H3					
Cromossoma	miRNA	Localização	Ref	Alt	Zigotia
1	hsa-mir-4253	23189715	G	C	het
1	hsa-mir-3118-1	142667306	C	T	het
1	hsa-mir-3118-2	143163772	G	C	het
1	hsa-mir-3118-2	143163775	A	G	het
1	hsa-mir-3118-2	143163790	T	C	het
1	hsa-mir-3118-2	143163799	G	A	het
2	hsa-mir-1302-3	114340663	G	A	het
2	hsa-mir-663b	133014552	C	T	het
2	hsa-mir-663b	133014559	G	T	het
2	hsa-mir-663b	133014580	C	T	het
2	hsa-mir-663b	133014592	G	A	het
2	hsa-mir-663b	133014602	G	C	het
2	hsa-mir-663b	133014603	G	T	het
2	hsa-mir-663b	133014612	A	C	het
2	hsa-mir-663b	133014614	C	T	het
2	hsa-mir-663b	133014617	A	G	het
2	hsa-mir-663b	133014619	C	T	het
2	hsa-mir-663b	133014621	A	C	het
2	hsa-mir-663b	133014633	T	C	het
2	hsa-mir-663b	133014640	C	T	het
3	hsa-mir-1324	75679930	A	C	het
3	hsa-mir-1324	75679937	G	A	het
3	hsa-mir-1324	75679938	T	C	het
3	hsa-mir-1324	75679945	C	T	het
3	hsa-mir-1324	75679958	C	T	het
3	hsa-mir-1324	75679965	C	A	het
3	hsa-mir-1324	75679973	A	G	het
3	hsa-mir-1324	75679975	C	T	het
3	hsa-mir-1324	75679988	A	G	het
3	hsa-mir-1324	75679990	G	A	het
3	hsa-mir-1324	75679997	C	T	het
3	hsa-mir-4273	75787458	G	A	het
3	hsa-mir-4273	75787464	C	T	het
3	hsa-mir-4273	75787476	C	T	het
3	hsa-mir-4273	75787486	C	T	het
4	hsa-mir-943	1988193	T	C	hom
5	hsa-mir-4277	1708983	G	A	hom
5	hsa-mir-3141	153975576	G	A	het
5	hsa-mir-146a	159912418	C	G	het
5	hsa-mir-585	168690612	A	G	het
5	hsa-mir-1229	179225324	G	A	het
6	hsa-mir-548u	57254955	G	A	het
6	hsa-mir-548u	57254986	G	A	het
6	hsa-mir-548u	57255001	C	A	het
8	hsa-mir-1206	129021179	G	A	hom
8	hsa-mir-1234	145625538	C	G	het
9	hsa-mir-3152	18573360	G	A	het
9	hsa-mir-1299	69002248	C	T	het
9	hsa-mir-1299	69002251	A	G	het
9	hsa-mir-1299	69002270	G	A	het
9	hsa-mir-1299	69002271	T	A	het
9	hsa-mir-1299	69002273	G	C	het
9	hsa-mir-1299	69002274	G	A	het
9	hsa-mir-1299	69002283	A	T	het

9	hsa-mir-1299	69002294	C	T	het
9	hsa-mir-1299	69002295	G	A	het
9	hsa-mir-1299	69002315	C	T	het
9	hsa-mir-1299	69002317	T	C	het
9	hsa-mir-204	73424994	T	G	het
10	hsa-mir-4293	14425204	T	A	het
10	hsa-mir-1307	105154089	A	G	het
10	hsa-mir-202	135061112	C	T	het
11	hsa-mir-1304	93466866	G	T	hom
12	hsa-mir-196a-2	54385599	C	T	hom
12	hsa-mir-618	81329536	A	C	het
14	hsa-mir-4308	55344901	C	T	het
14	hsa-mir-412	101531854	A	G	het
14	hsa-mir-4309	103006047	G	C	het
15	hsa-mir-3175	93447631	T	G	het
16	hsa-mir-1826	33965518	G	C	het
16	hsa-mir-1826	33965520	C	T	het
16	hsa-mir-1826	33965529	C	T	het
16	hsa-mir-1826	33965533	A	C	het
16	hsa-mir-1826	33965537	C	T	het
16	hsa-mir-1826	33965538	A	G	het
16	hsa-mir-1826	33965542	C	T	het
16	hsa-mir-1826	33965554	A	G	het
16	hsa-mir-182	33965560	T	A	het
16	hsa-mir-182	33965561	T	C	het
16	hsa-mir-182	33965570	G	A	het
16	hsa-mir-182	33965582	C	T	het
16	hsa-mir-182	33965589	G	A	het
16	hsa-mir-182	33965591	A	C	het
16	hsa-mir-1972-2	70064261	C	T	het
17	hsa-mir-423	28444183	A	C	het
17	hsa-mir-2117	41522213	T	A	hom
19	hsa-mir-27a	13947292	T	C	het
19	hsa-mir-320e	47212593	A	G	hom
19	hsa-mir-520f	54185492	G	A	het
20	hsa-mir-663	26188824	T	C	het
20	hsa-mir-663	26188912	A	C	het
20	hsa-mir-499	33578251	A	G	het
20	hsa-mir-3196	61870167	C	A	het
20	hsa-mir-4326	61918164	C	T	het
21	hsa-mir-3156-3	14778721	A	T	hom
X	hsa-mir-532	49767832	A	G	hom
X	hsa-mir-532	49767835	A	G	hom

H4					
Cromossoma	miRNA	Localização	Ref	Alt	Zigotia
1	hsa-mir-4253	23189715	G	C	het
1	hsa-mir-3118-1	142667306	C	T	het
1	hsa-mir-3118-2	143163772	G	C	het
1	hsa-mir-3118-2	143163775	A	G	het
1	hsa-mir-3118-2	143163790	T	C	het
1	hsa-mir-3118-2	143163799	G	A	het
2	hsa-mir-1302-3	114340663	G	A	het
2	hsa-mir-663b	133014552	C	T	het
2	hsa-mir-663b	133014559	G	T	het
2	hsa-mir-663b	133014580	C	T	het
2	hsa-mir-663b	133014592	G	A	het
2	hsa-mir-663b	133014602	G	C	het
2	hsa-mir-663b	133014603	G	T	het
2	hsa-mir-663b	133014612	A	C	het
2	hsa-mir-663b	133014614	C	T	het
2	hsa-mir-663b	133014617	A	G	het
2	hsa-mir-663b	133014619	C	T	het
2	hsa-mir-663b	133014621	A	C	het
2	hsa-mir-663b	133014633	T	C	het
2	hsa-mir-663b	133014640	C	T	het
3	hsa-mir-1324	75679930	A	C	het
3	hsa-mir-1324	75679937	G	A	het
3	hsa-mir-1324	75679938	T	C	het
3	hsa-mir-1324	75679945	C	T	het
3	hsa-mir-1324	75679958	C	T	het
3	hsa-mir-1324	75679965	C	A	het
3	hsa-mir-1324	75679973	A	G	het
3	hsa-mir-1324	75679975	C	T	het
3	hsa-mir-1324	75679988	A	G	het
3	hsa-mir-1324	75679990	G	A	het
3	hsa-mir-1324	75679997	C	T	het
3	hsa-mir-4273	75787458	G	A	het
3	hsa-mir-4273	75787464	C	T	het
3	hsa-mir-4273	75787476	C	T	het
3	hsa-mir-4273	75787486	C	T	het
4	hsa-mir-943	1988193	T	C	hom
5	hsa-mir-4277	1708983	G	A	hom
5	hsa-mir-3141	153975576	G	A	het
5	hsa-mir-146a	159912418	C	G	het
5	hsa-mir-585	168690612	A	G	het
5	hsa-mir-1229	179225324	G	A	het
6	hsa-mir-548u	57254955	G	A	het
6	hsa-mir-548u	57254986	G	A	het
6	hsa-mir-548u	57255001	C	A	het
8	hsa-mir-1206	129021179	G	A	hom
8	hsa-mir-1234	145625538	C	G	het
9	hsa-mir-3152	18573360	G	A	het
9	hsa-mir-1299	69002248	C	T	het
9	hsa-mir-1299	69002251	A	G	het
9	hsa-mir-1299	69002270	G	A	het
9	hsa-mir-1299	69002271	T	A	het
9	hsa-mir-1299	69002273	G	C	het

9	hsa-mir-1299	69002274	G	A	het
9	hsa-mir-1299	69002283	A	T	het
9	hsa-mir-1299	69002294	C	T	het
9	hsa-mir-1299	69002295	G	A	het
9	hsa-mir-1299	69002315	C	T	het
9	hsa-mir-1299	69002317	T	C	het
9	hsa-mir-204	73424994	T	G	het
10	hsa-mir-4293	14425204	T	A	het
10	hsa-mir-1307	105154089	A	G	het
10	hsa-mir-202	135061112	C	T	het
11	hsa-mir-1304	93466866	G	T	hom
12	hsa-mir-618	81329536	A	C	het
14	hsa-mir-4308	55344901	C	T	het
14	hsa-mir-412	101531854	A	G	het
14	hsa-mir-4309	103006047	G	C	het
15	hsa-mir-3175	93447631	T	G	het
16	hsa-mir-1826	33965518	G	C	het
16	hsa-mir-1826	33965520	C	T	het
16	hsa-mir-1826	33965529	C	T	het
16	hsa-mir-1826	33965533	A	C	het
16	hsa-mir-1826	33965537	C	T	het
16	hsa-mir-1826	33965538	A	G	het
16	hsa-mir-1826	33965542	C	T	het
16	hsa-mir-1826	33965554	A	G	het
16	hsa-mir-1826	33965560	T	A	het
16	hsa-mir-1826	33965561	T	C	het
16	hsa-mir-1826	33965570	G	A	het
16	hsa-mir-1826	33965582	C	T	het
16	hsa-mir-1826	33965589	G	A	het
16	hsa-mir-1826	33965591	A	C	het
16	hsa-mir-1972-2	70064261	C	T	het
17	hsa-mir-423	28444183	A	C	het
17	hsa-mir-2117	41522213	T	A	hom
19	hsa-mir-27a	13947292	T	C	het
19	hsa-mir-320e	47212593	A	G	hom
19	hsa-mir-520f	54185492	G	A	het
20	hsa-mir-663a	26188824	T	C	het
20	hsa-mir-663a	26188912	A	C	het
20	hsa-mir-499	33578251	A	G	het
20	hsa-mir-3196	61870167	C	A	het
20	hsa-mir-4326	61918164	C	T	het
21	hsa-mir-3156-3	14778721	A	T	hom
X	hsa-mir-532	49767832	A	G	hom
X	hsa-mir-532	49767835	A	G	hom